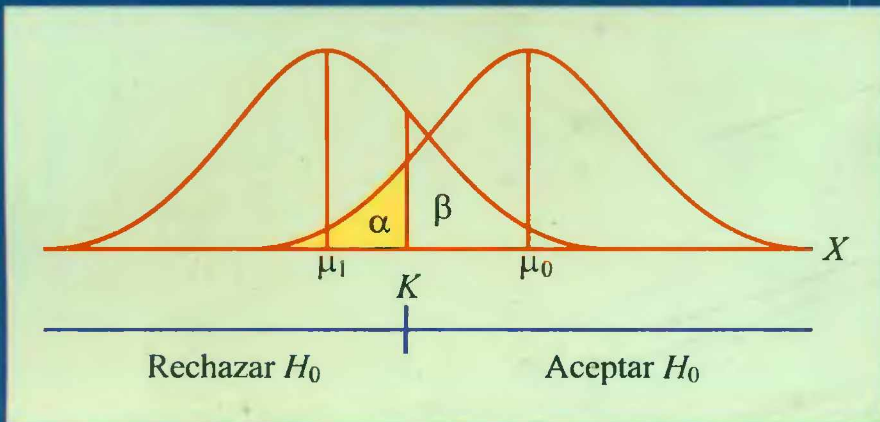
ESTADISTICA

Descriptiva e Inferencial



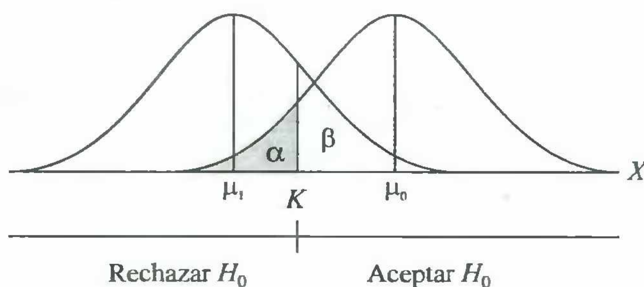
Aplicaciones

Quinta Edición

Manuel Córdova Zamora

ESTADISTICA

Descriptiva e Inferencial



Aplicaciones

Quinta edición

Mg. Manuel Córdova Zamora
email: mcordov@pucp.edu.pe

Profesor Principal T.C.
del Departamento de Ciencias de la
Pontificia Universidad Católica del Perú

ESTADISTICA : Descriptiva e Inferencial
Aplicaciones

Autor : Manuel Córdova Zamora

Derechos Reservados : Dec. Leg. 822
N° de Depósito Legal : 1501352000-1716
I.S.B.N. : 9972-813-05-3

Composición y diagramación : Manuel Córdova Zamora

Primera edición : Marzo 1995
Segunda edición : Agosto de 1996
Tercera Edición : Julio de 1997
Cuarta Edición : Junio del 2000
Quinta edición : Enero del 2003

Obra impresa en los talleres gráficos de :
Distribuidora, Imprenta, Editorial, Librería
MOSHERA S.R.L.
Jr. Tacna 2975 - San Martín de Porres - Lima - Perú
Telefax : 567-9299

Pedidos : Telf. : 534-0638

Impreso en el Perú - Printed in Perú

Presentación a la quinta edición

Esta publicación, como las cuatro ediciones anteriores, contiene los métodos de la experimentación estadística y la teoría de la probabilidad dando énfasis a las aplicaciones de manera que sirva a la formación básica que debe tener un estudiante de cualquier especialidad, y que sirva también, como fuente de consulta para profesionales de diversas disciplinas del saber.

Una de las grandes ventajas de este libro es la inclusión de numerosos ejemplos y ejercicios, algunos de ellos casos del mundo real, y que estimulan la curiosidad científica del lector permitiendo realizar una adecuada organización, presentación y análisis de los datos que se obtienen de la experimentación.

Se pretende que los ejercicios con datos experimentales conduzca al alumno a la formulación de modelos de probabilidad y de regresión. Para esto y otros problemas de aplicación, el autor le brinda el software MCEST cuyos resultados son compatibles con los de los paquetes de computo SPSS, ESTADÍSTICA. EXCEL MINITAB y otros.

El libro contiene 10 capítulos. En los tres primeros, se introduce los métodos descriptivos, como tabulación de datos en distribución de frecuencias, medidas de tendencia central, de dispersión asimetría y curtosis. El capitulo 4 está dedicado a una introducción al estudio de la regresión lineal simple en la forma descriptiva y los números índices. El capitulo 5 trata del cálculo de probabilidades. Los capítulos 6 y 7 constituyen una presentación básica de las variables aleatorias y distribuciones de probabilidades especiales. En los capítulos 8, 9 y 10 se hace una introducción a la inferencia estadística dando ideas básicas sobre estimación y pruebas de hipótesis

Estoy muy agradecido por la acogida que recibe esta publicación, y me motiva entonces a realizar una revisión permanente del texto ampliando sus aplicaciones. Quiero expresar también mi agradecimiento a la Pontificia Universidad Católica del Perú por permitirme realizar este trabajo fruto de mi experiencia en sus aulas. A mis alumnos de Estudios Generales Ciencias, Estudios Generales Letras, Ciencias Administrativas y Contabilidad. Así mismo expreso mi agradecimiento a todos los colegas de Lima y provincias por la aplicación del texto, el mismo que se ajusta a los programas de los cursos de Estadística General que forma parte de los planes de estudio de las Universidades e Institutos Superiores.

Lima, Enero del 2003

CONTENIDO

Capítulo 1

NOCIONES DE ESTADISTICA DESCRIPTIVA

1.1 Estadística

La palabra *estadística* se emplea en una gran variedad de formas. En plural se emplea como sinónimo de dato.

El trabajo estadístico o la *investigación estadística* es un proceso que pasa generalmente por las siguientes etapas:

- *Formulación del problema o la tarea.*
- *Diseño* del experimento.
- *Recopilación* de los datos.
- *Clasificación, tabulación y descripción* de resultados.
- *Generalización o inferencia.*

Definición. En este texto básico, definimos la estadística, como la ciencia que nos proporciona un conjunto de métodos, técnicas o procedimientos para:

- *recopilar,*
- *organizar (clasificar, agrupar),*
- *presentar, y*
- *analizar,*

datos con el fin de describirlos o de realizar generalizaciones válidas.

Se denomina ***estadística descriptiva,*** al conjunto de métodos estadísticos que se relacionan con el resumen y descripción de los datos, como tablas, gráficas, y el análisis mediante algunos cálculos.

Se denomina ***inferencia estadística*** al conjunto de métodos con los que se hacen la generalización o la inferencia sobre una población utilizando una muestra. La inferencia puede contener conclusiones que pueden no ser ciertas en forma absoluta, por lo que es necesario que éstas sean dadas con una medida de confiabilidad que es la **probabilidad**.

Estas dos partes de la estadística no son mutuamente excluyentes, ya que para utilizar los métodos de la inferencia estadística, se requiere conocer los métodos de la estadística descriptiva.

1.2 Población y muestra

Población

Definición. En forma general, en estadística; se denomina población, a un conjunto de elementos (que consiste de personas, objetos, etc.), que contienen una o más *características* observables de naturaleza cualitativa o cuantitativa que se pueden medir en ellos.

A cada elemento de una población se denomina **unidad elemental o unidad estadística**.

Por ejemplo, los empleados de una empresa en un día laborable, constituyen una población en la que cada empleado (unidad estadística), tiene muchas características a ser observadas, como por ejemplo: sexo, estado civil, lugar de procedencia, grado de instrucción, etc. (características cualitativas), o número de hijos, ingresos mensuales, etc.(características cuantitativas).

El resultado de medir una característica observable de una unidad elemental, se denomina **dato estadístico** o **valor observado** o simplemente **observación.**

Por otra parte, la población; viene definida por la tarea o investigación estadística a realizarse. Y como la medición o conteo de la característica especificada por la investigación se hace a cada unidad elemental, se puede considerar a la *población* como la *totalidad de valores posibles* de una característica particular especificada por la investigación estadística. En este sentido la población consiste de un conjunto de datos estadísticos que se reúnen de acuerdo con la formulación de una investigación estadística o con la definición de la población específica.

Parámetro: Se denomina parámetro a una medida descriptiva que resuma una característica de la población, tal como la media (μ) o la varianza (σ^2), calculada a partir de los datos observados de toda la población .

Tipos de población: Por el número de elementos que la componen, la población se clasifica en finita o infinita. La **población** es **finita** si tiene un número

finito de elementos. En caso contrario la **población** es **infinita**. En la práctica una población finita con un número grande de elementos se considera como una población infinita.

Muestra

Después de definir la investigación estadística a realizar, se debe decidir entre investigar toda la población o sólo una parte de ella. El primer procedimiento es denominado **censo** y el segundo es llamado **muestreo.**

Definición. Se denomina *muestra* a una parte de la población seleccionada de acuerdo con un plan o regla, con el fin de obtener información acerca de la población de la cual proviene.

La muestra debe ser seleccionada de manera que sea representativa de la población. Un método de selección de muestras representativas es al *azar simple*, esto es, cada elemento de la población tiene la misma posibilidad de ser seleccionada para la muestra.

Estadística o estadígrafo. Se denomina estadística a una medida descriptiva que resuma una característica de la muestra, tal como la media ($\overline{x}$) o la varianza (s^2) calculada a partir de los datos observados de una muestra aleatoria.

Es importante tener en cuenta, si el análisis estadístico se está haciendo con una muestra o con una población. En ambos casos las medidas descriptivas son las mismas. Para diferenciarlos, los parámetros de la población, se representan por letras griegas.

1.3 Variables estadísticas

La característica que se mida en las unidades elementales de una población definida por la tarea estadística, tiene diversos valores de naturaleza cualitativa o cuantitativa. Por ejemplo, la característica "sexo" tiene dos modalidades: hombre y mujer, la característica "peso en kilogramos" tiene infinitos valores.

Definición: Se denomina variable estadística a una característica definida en la población por la tarea o investigación estadística, que puede tomar dos o más valores (cualidades o números).

Se representa por una letra del alfabeto. Por ejemplo, en la población constituida por los empleados de la universidad, algunas variables estadísticas definidas en ésta población son:

X: "sexo". Valores: Masculino, Femenino

Y: "estado civil". Valores: Soltero, casado, viudo, divorciado

Z: "número de hijos", Valores: 0,1,2, etc.
W: "ingresos mensuales", Valores: Números reales positivos.



Figura 1.0. variable estadística

Si una variable se denota por *X*, entonces, sus valores observados en *n* unidades estadísticas se denotan por $x_1, x_2, ..., x_n$, conforme al orden en que se han obtenido. Este conjunto de *n* observaciones constituye una *muestra de tamaño n* obtenida de una población.

1.3.1 Escalas de medición

La asignación de valores a cada una de las unidades estadísticas mediante una variable, se hace siguiendo determinadas escalas de medición.

Definición. Se denomina escala de medición a un instrumento de medida, con el que se asigna valores (cualidades o números) a las unidades estadísticas para una variable definida.

El conocimiento de las escalas de medición es muy importante, pues cada una de ellas tienen métodos estadísticos específicos.

Las escalas de medición son de los siguientes tipos:

* *Nominal*
* *Ordinal*
* *De intervalos, y*
* *De razones.*

Escala nominal

Definición: Se tiene una escala nominal si dos o más valores de una variable, sólo permiten percibir las diferencias o semejanzas de las unidades estadísticas que se midan. Tales valores son como etiquetas que identifican a las unidades estadísticas y las hacen iguales o diferentes entre si.

Si se asignan números a estos valores cualitativos (modalidades), con estos no se pueden realizar operaciones aritméticas. Sólo son válidas las relaciones de igualdad (=) y no igualdad (≠).

Por ejemplo, la variable "sexo" asigna a las personas dos valores: "masculino" y "femenino" que son de escala nominal. Con los valores de esta variable **las personas están en una misma modalidad o en modalidades diferentes**. Si se asigna un "cero" al sexo "masculino" y un "uno" al sexo femenino, con estos números, no se pueden realizar operaciones aritméticas. Sólo se puede decir que el símbolo 0 es distinto al símbolo 1, pero no podemos decir que 1 es mayor que 0, o que 0 es menor que 1. Las variables estadísticas: "estado civil", "ideas religiosas", entre otras, tienen modalidades que son de escala nominal.

El método estadístico con datos obtenidos en escala nominal consiste básicamente en obtener el número o porcentaje de casos en cada modalidad y obtener la moda (valor de mayor frecuencia)

Escala ordinal

Definición: Una escala ordinal es una escala nominal donde los valores de la variable se pueden ordenar en forma ascendente (o descendente).

En una escala ordinal los valores o modalidades reflejan el orden de las unidades estadísticas. Si se asignan números a tales modalidades, con estos, no se pueden realizar operaciones aritméticas. Sólo son válidas las relaciones de igualdad (=), de no igualdad (≠) y de orden (≤).

Por ejemplo, la variable "estatus socioeconómico" con sus modalidades: clase baja, media y alta se mide, en escala ordinal. La variable "orden de mérito" cuyas modalidades son 1°, 2°, 3°, etc. mide las calificaciones de las unidades estadísticas en escala ordinal.

El método estadístico con datos obtenidos en escala ordinal consiste básicamente en obtener el número o porcentaje de casos en cada modalidad y obtener la moda, la mediana, los percentiles y el coeficiente de correlación por rangos.

Escala de intervalos

Definición: Una escala de intervalos es una escala ordinal con cuyos "valores" no sólo se pueden verificar igualdad , no igualdad y orden, si no también, se puede elegir una unidad de escala y comprobar *cuántas veces la diferencia entre dos valores es igual a la diferencia entre otros dos valores de la escala* (es decir, podemos comparar intervalos).

Esto es, si x_1, x_2 y x_3 son tres valores en la escala de intervalo, se verifica, por ejemplo, la relación:

$$x_3 - x_1 = c(x_2 - x_1) \quad \text{o} \quad \frac{x_3 - x_1}{x_2 - x_1} = c$$

donde c es una constante.

Esta relación se interpreta como que la escala de intervalos tiene un **"cero relativo"**. Este cero no significa ausencia total de la propiedad que se observa.

Se miden en escala de intervalos, por ejemplo, la temperatura (grados Celsius o Fahrenheit), el tiempo que se registra en nuestros calendarios, las calificaciones de una prueba de conocimientos o de aptitud. Estas mediciones tienen un cero elegido arbitrariamente, por ejemplo, el agua se congela a 0°C(=32°F). La calificación "cero" de un alumno en un examen de matemática básica, no quiere decir que no sabe nada de tal materia, pues con otra prueba más fácil podría tener otra calificación.

Con los valores de esta escala son válidas las relaciones de igualdad, de no igualdad y de orden. También, son válidas las operaciones de adición y sustracción entre los valores de la escala, y la multiplicación y división entre las diferencias de dos valores de la escala. Pero, no es válida la multiplicación y división entre los valores mismos de la escala.

Por **ejemplo**, si la variable X es "puntaje" obtenido en un examen calificado de 2 a 20 donde la unidad de medida es un punto a partir de 2 (cero relativo), entonces, se tiene una escala de intervalos. En efecto, si tres alumnos A, B, y C han obtenido los puntajes: 2, 4 y 16 respectivamente, no solo se verifican las relaciones : $2 \neq 4 \neq 16$ y $2 < 4 < 16$, sino también: $16 - 2 = 7(4 - 2)$; es decir la diferencia de los puntajes de C menos A es igual a siete veces la diferencia de los puntajes de B menos A. No es válida la división 16/4, pero si lo es, (16–2)/(4–2).

Propiedad: Una escala de intervalo x permanece invariante ante la transformación:

$$y = ax + b$$

donde a y b son constantes arbitrarias. Esto se debe a que son arbitrarios tanto el origen como la unidad de medida.

Por **ejemplo**, si la variable X tiene valores: 2, 4 y 16 que se miden en escala de intervalos, entonces, la transformación: $Y = 3X - 3$, que produce los valores respectivos 3, 9 y 45, **es la misma escala de intervalos**. Es decir, estos dos juegos de valores 2, 4, 16, y 3, 9, 45 miden las misma característica y verifican las mismas relaciones:

$$2 \neq 4 \neq 16,\ \ 2<4<16,\ \ 16-2=7(4-2),\ \text{y}$$

$$3 \neq 9 \neq 45,\ \ 3<9<45,\ \ 45-3=7(9-3)$$

En particular :
$$\frac{16-2}{4-2} = 7 = \frac{45-3}{9-3}$$

Escala de razón o cociente

Definición. La escala de razón es una escala de intervalo con cuyos valores además podemos comprobar cuántas veces un valor de la escala es igual a otro valor de la escala. Esto es, si x_1 y x_2 son dos valores en la escala de razón, se verifica la relación: $x_2 = cx_1$ o $\frac{x_2}{x_1} = c$,

donde c es una constante, y $x_1 \neq 0$.

La escala de razón *tiene un cero* absoluto (ausencia total de la característica que se observa). Con los números de esta escala son válidas las relaciones de igualdad, de no igualdad, de orden y todas las operaciones matemáticas. Los valores de esta escala se obtienen en general, por mediciones tipo conteo (discretos) o por mediciones tales como de longitud, peso, volumen, vida útil, etc. (continuos).

Por **ejemplo**, si la variable X, es la longitud (en metros) de un objeto, entonces, los valores de esta variable son de escala de razón. En efecto, si tres objetos A, B, y C miden 2, 4 y 16 metros, se pueden establecer las relaciones: $2\neq4\neq16$, $2 < 4 < 16$, $16 - 2 = 7(4 - 2)$, además, $4/2 = 2$, $16/2 = 8$, y $16/4 = 4$. Es decir, la longitud de B es el doble que la de A, el de C es 8 veces que la de A y el de C es 4 veces que la de B.

Propiedad: Una escala de razón x permanece invariante ante la transformación:

$$y = ax$$

donde a es una constante arbitraria.

Por **ejemplo**, si la variable X: tiene valores 4 y 16 medidos en escala de razón, entonces, la transformación $Y = (1/2)X$ que produce los valores respectivos 2 y 8 **es la misma escala de razón**. Es decir, estos dos juegos de valores 4, 16 y 2, 8, miden las misma característica y verifican las mismas relaciones, en particular:

$$\frac{16}{4} = 4 = \frac{8}{2}$$

Por dar **otro ejemplo**, si el ingreso mensual de una persona se expresa en soles (x) o en dólares (y), entonces se cumple la relación: $y = (1/3.5)x$, donde 3.5 es el tipo de cambio.

NOTA. La aplicación de métodos estadísticos cuantitativos requieren que la variable se mida por lo menos en escala de intervalos.

EJEMPLO 1.0

Las notas de un cierto curso se miden en una escala de intervalos de 0 a 20. Por razones prácticas se trata de expresar estas notas en la misma escala de intervalos pero de tal manera que el 20 se transforme en 100 y el 15 se transforme en 80, ¿en cuánto debe transformarse el 0?

SOLUCION

Sea ? el valor que corresponde al 0.Entonces, si 0, 15, 20 y ?, 80, 100 son dos juegos de valores en la misma escala de intervalos, entonces,

$$\frac{?-80}{80-100}=\frac{0-15}{15-20}\text{, de donde resulta } ?=20.$$

1.3.2 Clasificación de variables.

Las variables se clasifican en cualitativas y cuantitativas. Las variables cuantitativas se clasifican en discretas y continuas

Variable cualitativa, es la característica cuyos valores se expresan en escalas nominal u ordinal, por ejemplo, sexo , profesión ,estado civil, orden de méritos, etc.. Con sus valores, que son cualidades, no se pueden realizar operaciones aritméticas.

Variable cuantitativa, es la característica cuyos valores se expresan en escalas de intervalo o de razón, por ejemplo, temperatura, número de hijos, ingresos mensuales, tiempo de vida útil, etc.. Con sus valores, que son números, se pueden realizar operaciones aritméticas.

Las variables cuantitativas, a su vez, se clasifican en: *discretas, y continuas.*

Variable discreta, es aquella variable cuantitativa que puede tomar sólo ciertos valores en un intervalo considerado y no admite ningún valor entre dos valores consecutivos fijos. Generalmente, es una variable cuyos valores *se obtienen por conteo* (números naturales). Por ejemplo, una familia puede tener 0,1,2,...hijos, pero no algún valor intermedio.

Variable continua, es aquella variable cuantitativa que puede tomar cualquier valor en el intervalo considerado, por ejemplo, salario, tiempo, peso, volumen, longitud, etc..

La distinción entre variable discreta y continua es más teórica que real. Al utilizar los datos, la variable siempre resulta discreta, pues toda medición se expresa sólo en ciertas unidades realmente medibles, por decir, metros, decímetros, centímetros. Por ejemplo, es posible que una persona mida 1.6748m, o, 1.6752m, pero para fines prácticos, **redondeando** a dos decimales, se considera sólo 1.67m., o, 1.68m respectivamente.

Una variable continua puede pues tomar infinitos valores intermedios en un intervalo dado. Para fines prácticos los valores numéricos de las variables continuas siempre son valores **aproximados**.

1.4 Organización de los datos: Distribución de frecuencias.

Después de la recopilación de los datos, es necesario resumirlos y presentarlos en forma tal, que faciliten su comprensión y su posterior análisis y utilización. Para ello, se ordenan en ***cuadros numéricos*** y luego se representan en ***gráficos***.

Existen muchos paquetes de computo estadísticos para organizar datos. Uno de estos es el **paquete estadístico *MCEST*** creado por el autor de este texto.

Todo cuadro numérico debe tener:

-Un *título* adecuado para evitar confusiones y para expresar brevemente su contenido.

-La *fuente* de los datos, si no son datos propios.

-Las *unidades* en que se expresan los datos.

Los cuadros numéricos de una sola variable estadística se denominan ***distribución de frecuencias***.

En el procedimiento para construir distribuciones de frecuencias *nos referiremos a muestras*, mientras no se diga lo contrario.

1.4.1 Distribución de frecuencias: Variable cualitativa

Supongamos que en una muestra de n unidades estadísticas se observan k categorías o modalidades diferentes $C_1, C_2, ..., C_k$, de alguna variable cualitativa X. La tabulación de estos n datos, es la distribución de frecuencias por categorías del cuadro 1-1.

La **frecuencia absoluta** f_i, es el número de datos observados en cada categoría o modalidad. **La suma de todas las frecuencias absolutas es igual al total n** de datos observados.

La **frecuencia relativa** h_i se define en cada categoría por $h_i = f_i/n$. **La suma de todas las frecuencias relativas es igual a uno**.

La **frecuencia porcentaje** p_i se define en cada fila por $p_i = h_i \times 100\%$. **El total de las frecuencias porcentajes es igual a cien**.

Cuadro 1.1. *Distribución de frecuencias de variable cualitativa*

Categorías de la Variable X	Frecuencias Absolutas f_i	Frecuencias Relativas h_i	Frecuencias Porcentajes p_i
C_1 C_2 ... C_k	f_1 f_2 ... f_k	h_1 h_2 ... h_k	p_1 p_2 ... p_k
Total	n	1.00	100.00

Gráfica.

Existe una gran variedad de gráficas para la distribución de frecuencias de variable cualitativa, las más comunes son la *de barras* y la de *sectores circulares*.

En una **gráfica de barras** los datos de cada una de las modalidades C_i se representan por una barra rectangular vertical (u horizontal), cuya altura (o largo) es proporcional a su frecuencia. Las barras se dibujan dejando un espacio entre ellas.

Si la escala es nominal las categorías pueden ser colocadas en cualquier orden. Pero, **si el nivel es ordinal las categorías deben ir ordenadas.**

En una **gráfica circular,** los datos de cada categoría C_i se representan por un sector circular cuyo ángulo en el centro es igual a $h_i \times 360°$.

Si la gráfica por sectores circulares es tridimensional es denominada **de pastel**.

EJEMPLO 1.1.

En una encuesta de opinión acerca de las preferencias de una marca de bebida gaseosa por sus colores: Negro(N), Blanco(B), Rojo(R), 20 consumidores dieron las siguientes respuestas:

B, N, N, B, R, N, N, B, B, N,
B, N, N, R, B, N, B, R, B, N.

Construir la distribución de frecuencias. Graficar la distribución

SOLUCION.

La tabulación de estos datos, donde la variable cualitativa es X: Color de bebida gaseosa, es la distribución de frecuencias del cuadro 1.2.

La figura 1.1 es la representación gráfica por medio de barras de la distribución de personas por el color de su bebida gaseosa preferida.

Cuadro 1.2. Distribución de personas por su color preferido de una marca de bebida gaseosa.

Valores de X	Frecuencias Absolutas: f_i	Frecuencias Relativas: h_i	Frecuencias Porcentajes: p_i
Negro (N)	9	0.45	45
Blanco (B)	8	0.40	40
Rojo (R)	3	0.15	15
Total	20	1.00	100

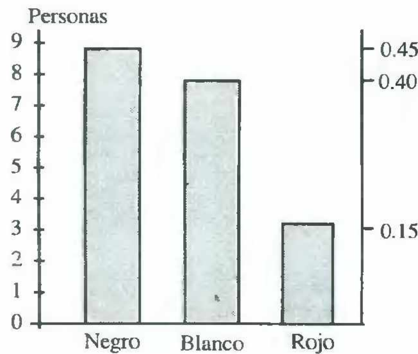


Fig. 1.1 Gráfica de barras

La figura 1.2 es la representación mediante gráfica de sectores circulares del cuadro 1.2. La frecuencia 45% es equivalente a $0.45 \times 360° = 162°$, la frecuencia 40% es equivalente a $0.40 \times 360° = 144°$, y la frecuencia 15% es equivalente a $0.15 \times 360° = 54°$

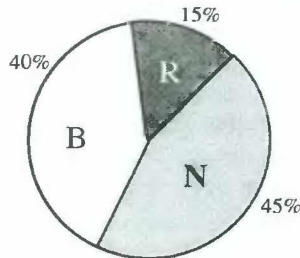


Fig. 1.2 Gráfica circular

1.4.2 Distribución de frecuencias: Variable cuantitativa discreta

Suponga que se han recolectado n valores de alguna variable discreta X. El procedimiento más simple de organizar estos n datos, consiste en ordenar estos valores numéricos en forma ascendente.

Si todos los n datos son distintos entre si, se obtendrá una distribución de frecuencias de n valores de la variable X, donde cada uno de estos valores tienen frecuencia absoluta igual a uno.

Si algunos valores se repiten, y si al terminar el ordenamiento se obtienen k ($k \leq n$) valores distintos de X, digamos, $x_1, x_2, ..., x_k$, con frecuencias absolutas respectivas $f_1, f_2, ..., f_k$, la distribución de frecuencias de estos n datos se resume en el cuadro 1.3 (observar que es similar al cuadro 1.1).

Cuadro 1.3. *Distribución de frecuencias de variable discreta*

Valores de La variable X	Frecuencias Absolutas f_i	Frecuencias Relativas h_i	Frecuencias Porcentajes p_i
x_1	f_1	h_1	p_1
x_2	f_2	h_2	p_2
...	...	...	...
x_k	f_k	h_k	p_k
Total	n	1.00	100.00

Las frecuencias absolutas relativas y porcentajes poseen , en el caso de variable discreta y continua, el mismo significado y propiedades, que en el caso de la variable cualitativa.

Cuando es grande el número de datos observados de una variable discreta, su organización es muy engorrosa. En este caso, para resumir los datos y poder calcular las medidas descriptivas, es conveniente seguir el método de organización de variable continua por intervalos que se describe en la sección 1.4.3 siguiente.

Gráfica

La representación gráfica más común de una distribución de frecuencias de variable discreta es del tipo **bastón** que consiste en trazar en cada valor distinto de la variable, segmentos de recta proporcionales a su frecuencia.

También, se pueden usar barras rectangulares para graficar una distribución de frecuencias de variable discreta.

EJEMPLO 1.2.

Ante la pregunta del número de hijos por familia (variable X) una muestra de 20 hogares, marcó las siguientes respuestas:

2. 1. 2. 4. 1. 3. 2. 3. 2. 0.

3, 2, 1, 3. 2, 3, 3, 1, 2, 4.

Construir la distribución de frecuencias de la variable X. Graficar.

SOLUCION.

Al ordenar estos datos en forma ascendente, se obtienen cinco valores distintos 0, 1, 2, 3, 4 que se repiten respectivamente 1, 4, 7, 6, 2 veces. La distribución de frecuencias de X se da en el cuadro 1.4.

Cuadro 1.4. Distribución de frecuencias del número de hijos por familia

Número de Hijos X_i	Frecuencias Absolutas f_i	Frecuencias relativas h_i	Frecuencias porcentajes p_i
0	1	0.05	5
1	4	0.20	20
2	7	0.35	35
3	6	0.30	30
4	2	0.10	10
Total	20	1.00	100

La gráfica de bastones es la figura 1.3

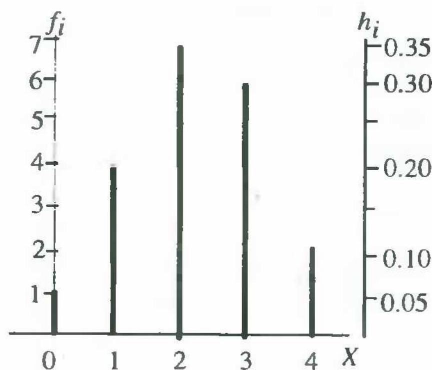


Figura 1.3. Gráfica de bastón

En la gráfica de bastones, figura 1.3, se indican las frecuencias absolutas y relativas en cada valor distinto de la variable.

1.4.3 Distribución de frecuencias por intervalos

La distribución de frecuencias por intervalos o clases se usa cuando la variable estadística es continua o cuando el número de valores distintos de una variable discreta es muy grande (más de 20 líneas en el monitor de una computadora)

Esta distribución se obtiene dividiendo el **rango** de variación de los datos en k intervalos y determinando el número de datos que contiene cada intervalo (Cuadro 1.5).

Cuadro 1.5. *Distribución de frecuencias por intervalos*

Intervalos I_i	Conteo	Frecuencias		
		Absolutas f_i	Relativas h_i	Porcentajes p_i
I_1	///...	f_1	h_1	p_1
I_2	///...	f_2	h_2	p_2
...	...	...	...	...
I_k	///...	f_k	h_k	p_k
Total		n	1.00	100.00

Para construir la distribución de frecuencias de intervalos hay varios procedimientos. En este texto se conviene y recomienda:

R1. Elegir no más de 20 intervalos ni menos de 5, ya que muchos intervalos pueden complicar innecesariamente los cálculos de las medidas descriptivas, y pocos intervalos podrían omitir características importantes de los datos. En este texto *elegimos todos los intervalos de igual amplitud A*.

R2. El número de intervalos elegido, debe dar una distribución de frecuencias *mono modal*, es decir, una distribución cuyas frecuencias van aumentando progresivamente hasta una frecuencia máxima y luego van disminuyendo también progresivamente.

Construcción de la distribución de frecuencias

Dados n valores de alguna variable cuantitativa X continua (o discreta con más de 20 valores distintos) uno de los métodos para construir la distribución de frecuencias es:

1. Determinar el rango (R) de variación de los datos que se define por

$$R = X_{\max} - X_{\min}$$

donde $X_{\max}$ es el dato máximo y $X_{\min}$ es el dato mínimo.

2. Determinar el número de intervalos, k, teniendo en cuenta la recomendación **R1**.

Un valor aproximado del número de intervalos, k, nos proporciona la *regla de Sturges*, donde,

$$k = 1 + 3.3\log(n), \quad n \geq 10,$$

redondeado el número al entero inmediato mayor.

Por ejemplo, si se tienen $n = 45$ datos sin decimales, entonces, $k = 1 + 3.3\log(45) = 6.4556$. Luego, k podrá elegirse como 6, 7, 8, o cualquier otro número entero, teniendo en cuenta las recomendaciones **R1** y **R2**.

Alternativamente se puede utilizar $k = \sqrt{n}$, donde $25 \leq n \leq 400$.

3. Determinar la amplitud A del intervalo, dividiendo el rango entre el número de intervalos. Esto es,

$$A = R/k.$$

Si la división $A = R/k$ no es exacta en el número de decimales de los datos, entonces, el número A se aproxima por *exceso* de manera que se cubra todo el rango, esto es, de manera que $kA \geq R$.

Si los datos son enteros, A es entero, si los datos tienen un decimal, A tiene un decimal, etc. Por ejemplo, si los datos tienen dos decimales y si $R/k = 5.3416$, se elige $A = 5.35$. (no 5.34).

4. Determinar los extremos de los intervalos de la siguiente manera:

$$\begin{aligned} I_1 &= [X_{min},\ X_{min} + A[\\ I_2 &= [X_{min} + A,\ X_{min} + 2A[\\ I_3 &= [X_{min} + 2A,\ X_{min} + 3A[\\ &\ldots \qquad \ldots \\ I_k &= [X_{min} + (k-1)A,\ X_{min} + kA] \end{aligned}$$

Observe que se cierra por la derecha el último intervalo. Esto se debe a que si la división R/k es exacta en el número de decimales de los datos, entonces, $X_{max} = X_{min} + kA$.

EJEMPLO 1.3.

Los ingresos quincenales en dólares (variable X) de 45 personas son:

63	89	36	49	56	64	59	35	78
43	53	70	57	62	43	68	62	26
64	72	52	51	62	60	71	61	55
59	60	67	57	67	61	67	51	81
53	64	76	44	73	56	62	63	60

Construir una distribución de frecuencias de 8 intervalos.

SOLUCION:

1) De los datos, se encuentra $X_{max} = 89$ y $X_{min} = 26$. El rango de variación de los datos es: $R = 89 - 26 = 63$

2) La amplitud del intervalo se elige a partir del valor

$$A = \frac{R}{k} = \frac{63}{8} = 7.875.$$

Como los datos son enteros, elegimos $A = 8$.

3) Los intervalos, el conteo y las frecuencias absolutas de los 45 ingresos quincenales se dan en el cuadro 1.6:

Cuadro 1.6. *Distribución de los ingresos de 45 personas*

Intervalos I_i	Conteo	Frecuencias Absoluta f_i	Relativa h_i	Porcentaje $p_i\%$
[26,34[	/	1	0.022	2.2
[34,42[	//	2	0.044	4.4
[42,50[	////	4	0.089	8.9
[50,58[	///// /////	10	0.222	22.2
[58,66[	///// ///// ///// /	16	0.356	35.6
[66,74[	///// ///	8	0.178	17.8
[74,82[	///	3	0.067	6.7
[82,90]	/	1	0.022	2.2
Total		45	1.000	100.0

NOTA. El **paquete estadístico *MCEST*** es una herramienta de trabajo muy útil, entre otros paquetes para construir distribuciones de frecuencia. El lector debería verificar, por ejemplo, esta salida.

Marca de clase

Definición: La **marca** de clase o marca del intervalo $I_i = [L_i, U_i[$ es el número m_i, que se define como el punto medio del intervalo. Esto es,

$$m_i = \frac{L_i + U_i}{2}$$

Por ejemplo, las marcas de clase de los intervalos del cuadro 1.6 son respectivamente 30, 38, 46, 54, 62, 70, 78 y 86.

La marca de clase es el número que *representa a todos los datos contenidos en el intervalo*. Por ejemplo, el intervalo [58,66[del cuadro 1.6 contiene 16 datos, ahora se deja sin efecto su valores reales y cada uno de ellos es representado por la marca del intervalo cuyo valor es 62.

Es evidente pues, que al representar los datos tabulados en un intervalo por sus correspondientes marcas de clase, se dejan sin efecto los valores recopilados, por lo que se pierde alguna información. Por tanto, el cálculo de las medidas descriptivas a partir de la distribución de frecuencias da *valores aproximados* de los valores exactos.

Gráfica de la distribución por intervalos

La distribución de frecuencias se representa gráficamente por medio de un *histograma*, o de un *polígono de frecuencias*. A partir del polígono de frecuencias se puede trazar la *curva de frecuencias.*

Histograma

Es una gráfica de barras rectangulares verticales juntas. La base de cada barra es proporcional a la amplitud del intervalo, y la altura es proporcional a su frecuencia (absoluta, o relativa, o porcentaje).

En el eje horizontal se colocan las escalas de la variable. En el eje vertical se colocan las escalas de las frecuencias.

Los números representativos de la escala de los intervalos son generalmente las marcas de clase de cada intervalo. Aunque también pueden colocarse los límites de los intervalos.

La figura 1.4 es el histograma con frecuencias absolutas de la distribución de frecuencias del cuadro 1.6.

Polígono de frecuencias

Es una gráfica poligonal cerrada , que se obtiene uniendo con segmentos de recta, los puntos que tienen proporcionalmente como abscisa a la marca de clase y como ordenada a la frecuencia respectiva. Se cierra en ambos extremos en las marcas de clase adyacentes de frecuencia cero.

La figura 1.5 es el polígono de frecuencias del cuadro 1.6.

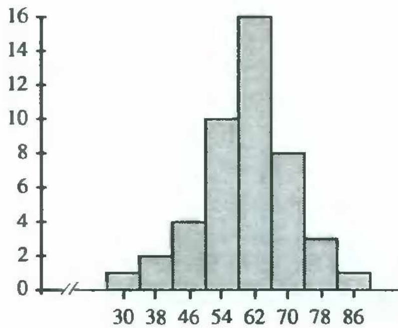


Fig. 1.4. Histograma

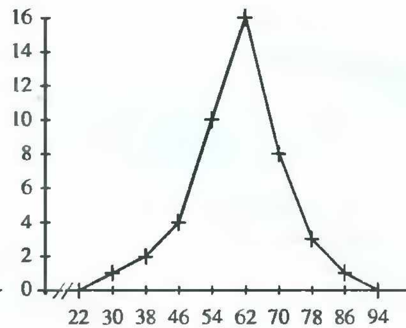


Fig. 1.5. Polígono

Curva de frecuencias

Una curva de frecuencias se obtiene del polígono de frecuencias "suavizando" sus puntos angulosos. En el proceso de suavización se recomienda tener en cuenta que la "porción" de área que se descarta deberá ser proporcional a la "porción" de área que se incluye en el interior de la gráfica.

La figura 1.6 es la curva de frecuencias absolutas de la distribución de frecuencias del cuadro 1.6.

La curva de frecuencias es importante por que representa realmente el tipo de población de la que se han obtenido los datos.

La curva de frecuencias es también llamada **modelo de la población**, y describe las características de la distribución de la población como: simetría, asimetría, tipos como: normal, bimodal, uniforme, etc..

Las curvas de frecuencias pueden pues tener una gran variedad de formas. Algunas de ellas son las siguientes:

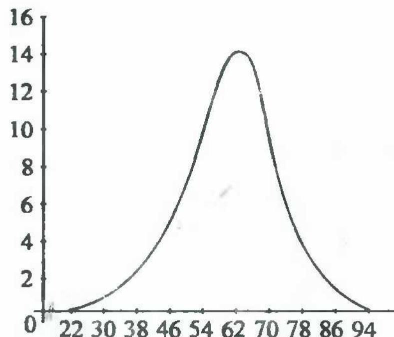


Fig. 1.6. Curva de frecuencias

Distribuciones Simétricas

Las curvas simétricas son de 3 tipos:

Normal o mesocúrtica (figura 1.7(a))

Platicúrtica (figura 1.7(b)), *y*

Leptocúrtica (figura 1.7(c)).

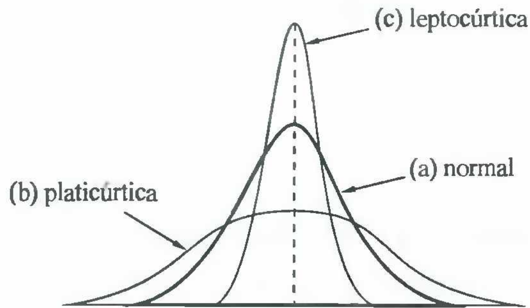


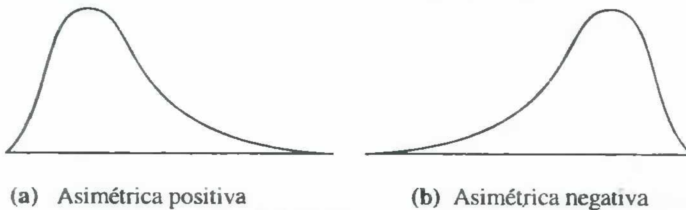
Fig. 1.7 Curvas simétricas: (a) normal, (b) platicúrtica, (c) leptocúrtica

Distribuciones Asimétricas

Las curvas asimétricas pueden ser de dos tipos:

Asimétricas positivas (o de cola al lado derecho), figura 1.8 (a).

Asimetría negativa(o de cola a la izquierda), fig. 1.8(b).



(a) Asimétrica positiva **(b)** Asimétrica negativa

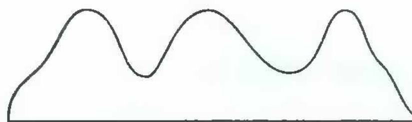
Fig. 1.8 Curvas asimétricas

Distribución multimodal

Una curva de frecuencias es bimodal si tiene dos frecuencias máximas, como la figura 1.9(a). Es trimodal si tiene tres frecuencias máximas, como la figura 1.9(b), etc..



(**a**) Curva bimodal



(**b**) Curva trimodal

Fig. 1.9 Curvas multimodales

1.4.4 Distribución de frecuencias acumuladas

A) Con **variables cualitativas** a nivel nominal *no tienen ningún significado las frecuencias acumuladas.*

B) **Si la variable es discreta** y la distribución de frecuencias es de la forma dato y frecuencia, entonces, las frecuencias (absolutas, relativas y porcentajes) se pueden acumular en la forma *menor o igual que* un valor determinado de la variable correspondiente a cada fila.

Si la **variable discreta** X tiene valores distintos $x_1, x_2, ..., x_k$ con frecuencias absolutas respectivas $f_1, f_2, ..., f_k$, entonces, la **frecuencia absoluta acumulada,** F_i, del valor x_i es la suma de las frecuencias absolutas de los valores **menores o iguales** a x_i, esto es,

$$F_i = \sum_{j=1}^{i} f_j, \quad i = 1,2,...,k.$$

La **frecuencia relativa acumulada** H_i de x_i, se define por:

$$H_i = \sum_{j=1}^{i} h_j \quad \text{o} \quad H_i = \frac{F_i}{n}, \qquad i = 1,2,3,...,k.$$

La **frecuencia porcentaje acumulada** P_i de x_i se define por:

$$P_i = H_i \times 100\%, \quad i = 1,2,...,k.$$

Por **ejemplo**, en el cuadro 1.7, se muestran las frecuencias acumuladas del número de hijos de 20 familias del cuadro 1.4. Aquí,

$$F_3 = \sum_{j=1}^{3} f_j = f_1 + f_2 + f_3 = 1 + 4 + 7 = 12.$$

El valor 12 de F_3 significa que existen 12 familias que tienen 2 hijos o menos, o que el 60% de las familias tienen 2 hijos o menos.

Cuadro 1.7. *Distribución de frecuencias acumuladas del número de hijos por familia*

# de hijos	Frecuencias			Frecuencias Acumuladas		
	Absoluta	Relativa.	Porcent.	Absoluta	Relativa.	Porcent.
X_i	f_i	h_i	p_i %	F_i	H_i	P_i %
0	1	0.05	5	1	0.05	5
1	4	0.20	20	5	0.25	25
2	7	0.35	35	12	0.60	60
3	6	0.30	30	18	0.90	90
4	2	0.10	10	20	1.00	100
Total	20	1.0000	100			

Gráfica.

La figuras 1.10(a) y 1.10(b) representan la distribución de frecuencias acumuladas absolutas y relativas respectivamente, "menor o igual que" del número de hijos por familia del cuadro 1.7.

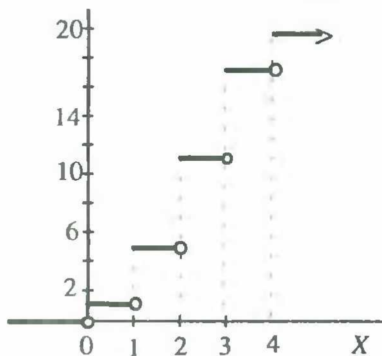


Fig. 1.10 (a)

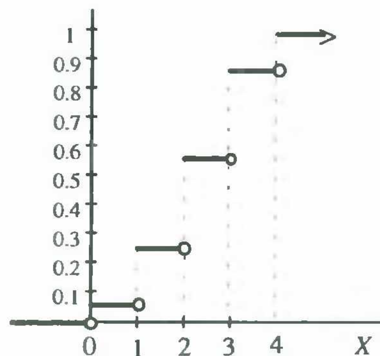


Fig. 1.10 (b)

Fig. 1.10. Gráfica de frecuencias acumuladas de variable discreta

NOTA. (Función de distribución acumulada). Las gráficas 1.10 son en realidad de una función denominada *función de distribución acumulada* (FDA) o simplemente *función de distribución.* Una definición más rigurosa de esta función se hará más adelante en el capítulo 6.

Por ejemplo, la figura 1.10.b donde las frecuencias son relativas, corresponde, a la función de distribución acumulada que sigue:

$$H(x)=\begin{cases}0, & \text{si} \quad x<0\\ 0.05, & \text{si } 0\le x<1\\ 0.25, & \text{si } 1\le x<2\\ 0.60, & \text{si } 2\le x<3\\ 0.90, & \text{si } 3\le x<4\\ 1.00, & \text{si} \quad x\ge 4\end{cases}$$

En general, si la variable discreta X tiene valores distintos $x_1, x_2, ..., x_k$ con frecuencias relativas respectivas $h_1, h_2, ..., h_k$, entonces, su FDA **menor o igual que** x para todo x real, se define por:

$$H(x)=\begin{cases}0, & \text{si } x<x_1\\ \sum_{j=1}^{i} h_j, & \text{si } x_i\le x<x_{i+1},\ i=1,2,...k-1\\ 1, & \text{si } x\ge x_k\end{cases}$$

C) Si la **distribución de frecuencias es de intervalos**, la frecuencia acumulada de cada intervalo es la **suma de las frecuencias** (absolutas, relativas o porcentajes) **hasta ese intervalo**.

En el cuadro 1.8 se presentan las frecuencias acumuladas de los ingresos quincenales de 45 personas que corresponden al cuadro 1.6.
En este cuadro,

$F_5 = 33$ indica que 33 ingresos son menores que $66

Gráfica.

La más usada es el *polígono de frecuencias acumuladas*, conocida también como ***ojiva***. La ojiva, se obtiene uniendo con segmentos de recta, los puntos cuya abscisa es proporcional al límite superior U_i de cada intervalo y cuya ordenada es proporcional a la frecuencia acumulada respectiva (absoluta, relativa o porcentaje).

Cuadro 1.8. *Distribución de frecuencias acumuladas de ingresos*

Intervalos	Frecuencias			Frecuencias Acumuladas		
	Absolut.	Relativa.	Porcent.	Absolut.	Relativa.	Porcent.
I_i	f_i	h_i	p_i %	F_i	H_i	P_i %
[26,34[	1	0.022	2.2	1	0.022	2.2
[34,42[	2	0.044	4.4	3	0.066	6.6
[42,50[	4	0.089	8.9	7	0.155	15.5
[50.58[	10	0.222	22.2	17	0.377	37.7
[58.66[	16	0.356	35.6	33	0.733	73.3
[66,74[	8	0.178	17.8	41	0.911	91.1
[74,82[	3	0.067	6.7	44	0.978	97.8
[82,90]	1	0.022	2.2	45	1.000	100.0
Total	45	1.000	100.0			

La figura 1.11 representa la distribución de frecuencias acumuladas de los ingresos quincenales de la muestra de 45 personas del cuadro 1.8.

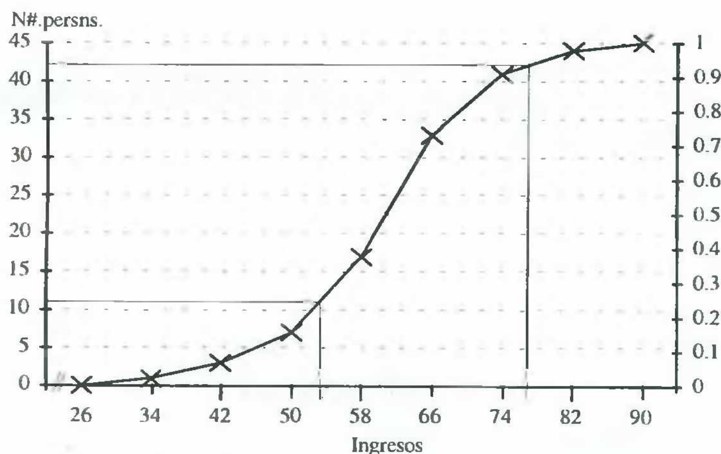


Fig. 1.11. Ojiva de la distribución de los ingresos de 45 personas

Aplicaciones de la Ojiva

Con la ojiva se pueden resolver dos tipos de problemas

1) Calcular el número (o porcentaje) de observaciones que corresponden a un intervalo determinado de la variable. Por **ejemplo**, en la figura 1.11, aproximadamente 16 personas (33–17 en el eje vertical) tienen ingresos entre 58 y 66 dólares quincenales (en el eje horizontal).

2) Determinar **cuantiles**.
   Se denominan **cuantiles** a los valores de la variable que dividen a la distribución de los datos en 2, 4, 10 o 100 partes iguales.

   Mediana. Es el valor Me de la variable que divide a la distribución en dos partes iguales.

   Cuartil. Es cada uno de los tres valores Q_1, Q_2, Q_3 que divide a la distribución de los datos en 4 partes iguales. El cuartil Q_2 es igual a Me.

   Percentil (o centil). Es cada uno de los 99 valores $P_1,..., P_{25},..., P_{50},..., P_{75},..., P_{99}$ que divide a la distribución de los datos en 100 partes iguales. El percentil $P_{25} = Q_1$, $P_{50} = Q_2$ y $P_{75} = Q_3$.

NOTA. Un dato observado está en el ***cuarto superior*** si es mayor que el tercer cuartil. Está en el ***cuarto inferior*** si es menor que el primer cuartil.

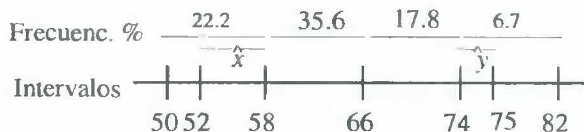
EJEMPLO 1.4.
En la distribución de los ingresos quincenales de 45 personas,
a) Determine el porcentaje aproximado de personas que tienen ingresos entre \$52 y 75\$.
b) Calcular el valor de los ingresos que divide a la distribución en dos partes iguales.
c) Calcular el percentil 25.

SOLUCION.

a) De la distribución de frecuencias se tiene que el porcentaje de ingresos quincenales comprendidos entre \$52 y \$75 es igual a:

$$x + 35.6 + 17.8 + y$$

donde x es el porcentaje comprendido entre \$52 y \$58, e y es el porcentaje comprendido entre \$74 y \$75, como se muestra en la figura que sigue.



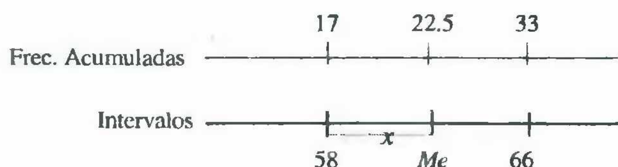
Los valores de x e y se calculan por **interpolación lineal** de la siguiente manera (Proviene de la semejanza de triángulos entre intervalos y frecuencias en la ojiva):

$$\frac{x}{58-52} = \frac{22.2}{58-50}, \text{ de donde resulta } x = 16.65$$

$$\frac{y}{75-74}=\frac{6.7}{82-74}, \text{ de donde resulta } y=0.8375$$

Luego el porcentaje aproximado de personas que tienen ingresos entre \$52 y 75\$ es igual a: $16.65+35.6+17.8+0.8375=70.89$.

b) Mediante la ojiva se ubica el intervalo que contiene a *Me* (mediana) y que corresponde al 50% (o *n*/2) de los datos.
En la distribución, $n/2=22.5$, entonces, $Me \in [58, 66[$.
Luego, $Me=58+x$, donde el valor de *x* se calcula de la figura que sigue:



$$\frac{x}{22.5-17}=\frac{66-58}{33-17}, \text{ de donde resulta } x=\frac{5.5}{16}\times 8=2.75$$

Luego, el valor que divide a la distribución en dos partes iguales es:

$$Me=58+2.75=60.75\$.$$

c) Mediante la ojiva se ubica el intervalo que contiene a P_{25} y que corresponde al 25% inferior (o $0.25n$) de los datos. En la distribución, $0.25\times n=11.25$, entonces, $P_{25} \in [50, 58[$. Además, $P_{25}=50+x$, el valor de *x* se calcula de:

$$\frac{x}{11.25-7}=\frac{58-50}{17-7}, \text{ de donde resulta: } x=\frac{4.25}{10}\times 8=3.4$$

Luego, el percentil 25 es el número: $P_{25}=50+3.4=53.4$ \$.

NOTA. El lector debería verificar que el valor del percentil P_k, (donde *k* es el percentil $k=1,\ldots,25,\ldots,50,\ldots,75,\ldots,99$), ubicado en el intervalo I_i es igual a:

$$P_k=L_{P_k}+\left[\frac{n(k/100)-F_{i-1}}{f_i}\right]A,$$

donde: L_{P_k} es el límite inferior del intervalo que contiene a P_k.

f_i es la frecuencia absoluta del intervalo que contiene a P_k.

F_{i-1} es la frecuencia acumulada absoluta del intervalo I_{i-1}

A es la amplitud del intervalo.

EJEMPLO 1.5

Calcular los percentiles 10, 75, y 90 para los 45 ingresos quincenales tabulados en la distribución de frecuencias del ejemplo 1.3 .

SOLUCION.

$$P_{10} = L_{P_K} + \left[\frac{n(10/100) - F_{i-1}}{f_i}\right]A = 42 + \left[\frac{4.5 - 3}{4}\right] \times 8 = 42 + 3 = 45.$$

$$P_{75} = Q_3 = L_{P_K} + \left[\frac{n(75/100) - F_{i-1}}{f_i}\right]A = 66 + \left[\frac{33.75 - 33}{8}\right] \times 8 = 66.75.$$

$$P_{90} = L_{P_k} + \left[\frac{n(90/100) - F_{i-1}}{f_i}\right]A = 66 + \left[\frac{40.5 - 33}{8}\right]8 = 66 + 7.5 = 73.5.$$

NOTA. (Percentiles en datos no tabulados). Para n datos **no tabulados** (pero ordenados) $x_1, x_2, ..., x_n$, el valor del percentil P_k es:

$$P_k = (x_j + x_{j+1})/2 , \quad \text{si} \quad g = 0, \quad \text{y} \quad P_k = x_{j+1}, \quad \text{si} \quad g > 0,$$

donde: $0.k \times n = j + g$, k = rango percentil, j = parte entera , g = parte fraccional.

EJEMPLO 1.6

Determinar los cuartiles P_{25}, P_{50}, P_{75}, para los siguientes datos:

1, 12, 4, 13, 8, 9, 10, 2, 7

SOLUCION

Al ordenar los datos, resultan: 1, 2, 4, 7, 8, 9, 10, 12, 13.

Para P_{25}, $0.25 \times n = 0.25 \times 9 = 2.25 = 2 + 0.25$, entonces, $j = 2$, $g = 0.25$.

Luego, $P_{25} = x_{j+1} = x_3 = 4$

Para P_{50}, $0.50 \times n = 0.50 \times 9 = 4.5 = 4 + 0.5$, entonces, $j = 4$, $g = 0.5$.

Luego. $P_{50} = x_{j+1} = x_5 = 8$

Para P_{75}, $0.75 \times n = 0.75 \times 9 = 6.75 = 6 + 0.75$, entonces, $j = 6$, $g = 0.75$.

Luego, $P_{75} = x_{j+1} = x_7 = 9$

EJEMPLO 1.7

Los ingresos mensuales de una muestra de n pequeños comerciantes se tabularon en una distribución de frecuencias simétrica de 5 intervalos de igual amplitud resultando: Ingreso mínimo \$125, marca de clase del cuarto intervalo $m_4 = \$300$. Si el 8% de los ingresos son menores que \$165 y el 70% de los ingresos son menores a \$275, ¿qué porcentaje de ingresos son superiores a \$285?.

SOLUCION.

Intervalos

$$X_{min} = 125, \ m_4 = \frac{(X_{min} + 3A) + (X_{min} + 4A)}{2} = 300, \quad \text{entonces}, \ A = 50$$

Luego, los intervalos son:

$$[125, 175[\ , [175, 225[\ , [225, 275[\ , [275, 325[\ , [325, 375].$$

Frecuencias

Si la distribución de frecuencias es simétrica, entonces, $h_1 = h_5$, $h_2 = h_4$.

Si el 70% de los ingresos son menores que $275, entonces:

$$h_1 + h_2 + h_3 = 0.7.$$

De $2h_1 + 2h_2 + h_3 = 1$, se tiene $h_1 + h_2 = 0.3$. Luego, $h_3 = 0.4$.

Además, si el 8% de los ingresos son menores que $165, entonces, $\frac{h_1}{50} = \frac{0.08}{40}$,

de donde resulta $h_1 = 0.10$. Finalmente $h_2 = 0.2$.

Luego, las frecuencias relativas son: 0.1, 0.2, 0.4, 0.2, 0.1.

Los ingresos superiores a $285 son: $\frac{4}{5}(0.2) + 0.1 = 0.26$ o 26%.

1.5 Otras gráficas estadísticas

En algunos casos, el total en cada modalidad de una variable, puede estar compuesto de varias partes, como por ejemplo, en el cuadro 1.9 la modalidad "año" tiene las componentes "hombres", "mujeres", y "total". El tipo de gráfico para este cuadro depende de lo que se quiere resaltar.

Cuadro 1.9. Población (en miles) de una ciudad de 1975 a 1990

Año	Hombres	Mujeres	Total
1975	8	17	25
1980	12	20	32
1985	10	30	40
1990	18	27	45

Gráfica de Línea

Si se quiere resaltar variaciones de los datos a través del tiempo, se utiliza una gráfica de líneas, la misma que se obtiene uniendo con segmentos, puntos de la forma: (tiempo, frecuencias) en cada modalidad y para una o más de las partes componentes. Por ejemplo, la figura 1.12 representa la población total del cuadro 1.9 desde 1975 a 1990.

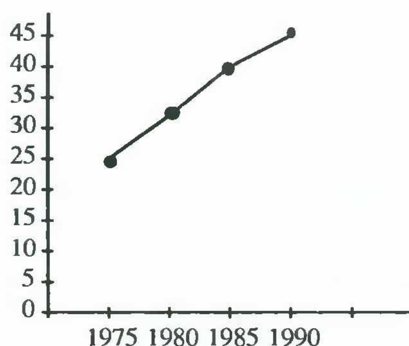


Fig. 1.12. Población total de 1975 a 1990

Gráfica de barras agrupadas

Si se trata de comparar solamente las componentes o las frecuencias en cada modalidad, se puede utilizar un gráfico de barras agrupadas. En cada modalidad se trazan tantas barras adjuntas como componentes hay. Por ejemplo, la figura 1.13 representa las frecuencias de cada componente en cada modalidad del cuadro 1.9.

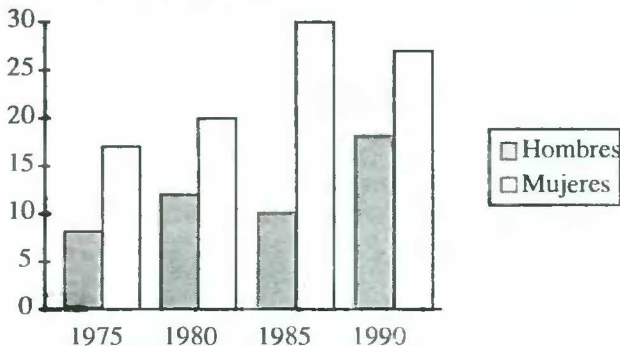


Fig. 1.13. Población de una ciudad de 1975 a 1990

Gráfica de barras componentes

a) Si se quiere resaltar a la vez el total y las frecuencias de cada componente en cada modalidad, entonces, conviene utilizar un **gráfico de barras componentes** como el de la figura 1.14. En cada modalidad se traza una barra cuyo largo es proporcional al total de sus datos. La gráfica 1.14 de barras componentes del cuadro 1.9 resume la variación de la población de una ciudad desde 1975 hasta 1990, resaltando el total y los parciales en cada modalidad.

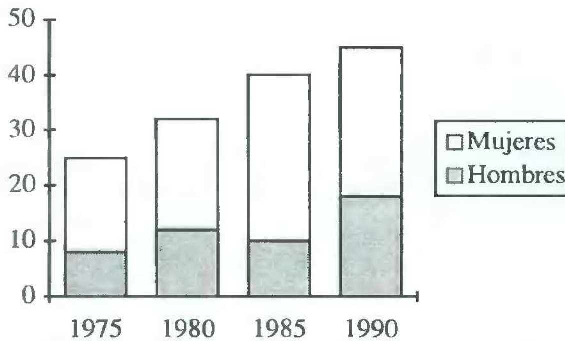


Fig. 1.14. Población de una ciudad de 1975 a 1990

b) Si se trata de destacar la importancia **relativa** de sus componentes, se puede utilizar un gráfico, como la figura 1.15, donde todas las barras son de igual longitud y equivalentes al 100% en cada categoría.

El cuadro 1.10, tiene los mismos datos del cuadro 1.9, sólo que ahora se consideran los porcentajes o valores relativos, en vez de los valores absolutos.

Cuadro 1.10. Población (en %) de una ciudad de 1975 a 1990

Año	Hombres	Mujeres	Total
1975	32.0	68.0	100
1980	37.5	62.5	100
1985	25.0	75.0	100
1990	40.0	60.0	100

La proporción de cada componente respecto al total en cada categoría, se representa en la figura 1.15.

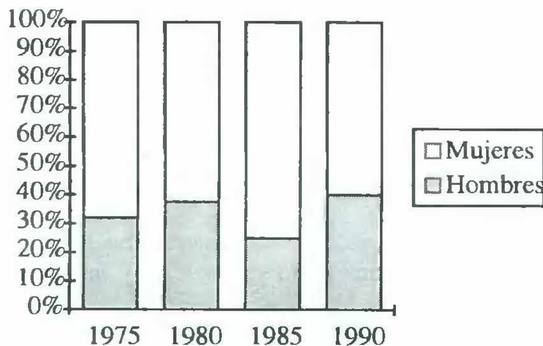


Fig. 1.15. Población de una ciudad de 1975 a 1990 en porcentajes

Cuando se utilizan figuras de igual tamaño para reflejar la característica que se quiere representar, al gráfico estadístico, se denomina **pictografía.** En una pictografía el número de figuras en cada categoría o modalidad es proporcional a la frecuencia absoluta respectiva.

Existe otra gran variedad de gráficas o diagramas para mostrar datos ó para mostrar relaciones entre varios grupos de datos. Aquí la imaginación del dibujante juega un papel muy importante.

EJERCICIOS

Variables y Escalas

1. Cierta variable asigna a las unidades estadísticas E_1 y E_2 de una población los valores 5 y 20 respectivamente en una escala dada. ¿Qué puede decir acerca de E_1 y E_2 si la escala usada es:

a) nominal?, b) ordinal?, c) de razón?.

Rp. a) $E_1 \neq E_2$, b) $E_1 < E_2$, c) 20/5=4. la medida de E_2 es igual a 4 veces a la de E_1

2. Cierta variable asigna los valores 1, 4 y 9 a las unidades estadísticas E_1, E_2, E_3 respectivamente en una escala de intervalos. Si en la misma escala se asigna 1 a E_1 y −8 a E_2 , ¿qué valor se le asigna a E_3?. Rp. −23.

3. Al medir cierta característica en una población a las unidades estadísticas E_1, E_2 y E_3 se les asigna los valores 2, 5 y 17 respectivamente usando una escala A. En cambio, usando una escala B, se asignan los valores 5 y 29 a E_2 y E_3 respectivamente.

a) ¿Podría afirmarse que A y B son la misma escala de razón?

b) ¿Qué podría afirmar sobre el valor de E_1 usando la escala B, si se sabe que ambas escalas son nominales? , son ordinales?, son la misma de intervalo?.

Rp. a) No, 17/5≠29/5. b) Nominal: cualquiera, ≠5 y ≠29, ordinal: <5, intervalo: −1.

4. Sean $x_1 = -1$, $x_2 = 0$, $x_3 = 1$. mediciones de una variable X a tres elementos de una población en una determinada escala. Suponga que son válidas las transformaciones:

$$\text{i) } Y = 2X - 5, \quad \text{ii) } Y = X^2 + 3$$

a) Si la escala es nominal. ¿están las mediciones transformadas también en escala nominal?
b) Si la escala es ordinal, ¿están las mediciones transformadas también en escala ordinal?
c) Si la escala es de intervalos, ¿están las mediciones transformadas en la misma escala de intervalos?

Rp. a) i) Si, ii) No, b) i) Si, ii) No, c) i) Si, ii) No.

5. Sean $x_1 = 0$ e $y_1 = 32$ dos valores asignados al mismo elemento para medir la temperatura, y, $x_2 = 100$ e $y_2 = 212$ dos valores asignados a la temperatura de otro elemento. Si los valores X (grados centígrados o Celsius) e Y (grados Fahrenheit) están en escala de intervalos, hallar la relación entre X e Y

$$Y = aX + b$$

Rp. a=9/5, b=32.

6. Al medir cierta característica en una población, las escalas A y B asignan valores x e y a un mismo elemento. Si la relación entre los valores es

$$y = \frac{x}{3} + 4$$

¿Son ambas escalas A y B, a) de intervalos?, b) de razón?.

Rp. a) Si A es de intervalos, B lo es también. b) Si A es de razón. B no lo es.

7. Clasifique las variables e indique el tipo de escala en que están medidas las siguientes características

- Profesión
- Nacionalidad
- Grado de instrucción
- Número de hijos
- Número de teléfono
- Dirección
- Año de nacimiento
- Edad
- Estado civil
- Ingreso mensual familiar promedio
- Número de DNI

Distribuciones de frecuencias

8. Al investigar el nivel socioeconómico en los valores: Bajo(B). medio (M), alto(A), 20 familias dieron las siguientes respuestas:

M, B, B, M, A, B, B, M, M, B, M, B, B, A, M, B, M, A, M, B.

Construir la distribución de frecuencias y trazar su gráfica.

9. Se revisaron 20 lotes de 48 artículos cada uno y se encontró el siguiente número de artículos defectuosos por lote:

3, 2, 5, 0, 1, 3, 2, 1, 0, 1, 3, 4, 2. 4. 4. 3, 4, 3. 2, 3.

Construir la distribución de frecuencias relativas y frecuencias relativas acumuladas. Graficar. ¿Qué porcentaje de lotes tienen dos o más pero menos de 4 artículos defectuosos?. Rp.50%.

10. Determinar los intervalos de la distribución de frecuencias en cada uno de los siguientes casos:

a) Datos enteros, $X_{min} = 10$, $X_{max} = 50$, y $k = 8$ intervalos.

b) Datos con dos decimales, $X_{min} = 2.55$, $X_{max} = 3.86$, y $k = 7$.

c) Datos con tres decimales, $X_{min} = 0.282$, $X_{max} = 0.655$, y $k = 6$.

Rp. a) A=5, b) A=0.19, c) A=0.063

11. La inversión anual, en miles de dólares, de una muestra de 40 pequeñas empresas fueron:

31	17	27	20	28	10	34	25	4	24
15	39	18	30	41	26	12	46	18	23
36	19	29	37	33	27	27	24	26	31
25	28	33	28	22	23	31	29	35	21

a) Construir una distribución de frecuencias de 7 intervalos de clase.

b) Determinar el porcentaje de empresas con una inversión entre 14 mil y 20 mil dólares. Rp. a) A = 6. Frecuencias: 1, 3, 6, 12, 11, 5, 2, b) 12.5%.

12. Se registra el tiempo en minutos que utilizan 30 alumnos para ejecutar una tarea, resultando los siguientes:

21.3	15.8	18.4	22.7	19.6	15.8	26.4	17.3	11.2	23.9
26.8	22.7	18.0	20.5	11.0	18.5	23.0	24.6	20.1	16.2
08.3	21.9	12.3	22.3	13.4	17.9	12.2	13.4	15.1	19.1

a) Construir una distribución de frecuencias de 6 intervalos de igual amplitud y a partir de ésta

b) Calcular el tiempo debajo del cual se encuentran el 25% de las tareas.

Rp. A = 3.1, frecuencias: 3, 4, 5, 8, 6, 4. b) 14.81

13. Las notas del examen parcial de matemática dieron la siguiente distribución de frecuencias

a) Completar la distribución de frecuencias.

b) Graficar la ojiva de porcentajes.

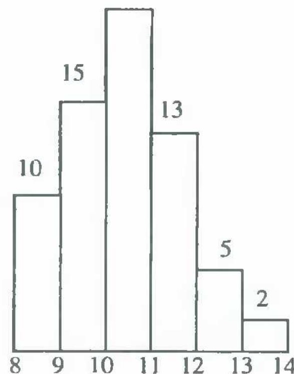
c) ¿Qué porcentaje de las notas se encuentran aproximadamente en el intervalo:[8, 14].

Intervalo	Marca de clase	Frecuencia Relativa	Frec. relativa acumulada
[, [		0.15	
[6 , [			0.45
[, [			0.70
[, [	13.5		
[,]		0.10	

Rp. a) $Xmin$=3, A=3, frec.: 0.15, 0.30, 0.25, 0.20, 0.10, c) 48.3%.

14. La distribución de los tiempos, en minutos, que utilizaron 65 personas para realizar una prueba de aptitud aparece representada en el siguiente histograma. ¿Qué porcentaje de las personas emplearon entre 9 y 11.5 minutos ?.

Rp. (15+20+6.5)/65=0.6385.



15. En una compañía, el sueldo mínimo y máximo de 200 empleados es de $150 y $300 respectivamente. Tales sueldos se tabulan en una distribución de frecuencias de 5 intervalos de igual amplitud. Si se sabe que 20 empleados ganan al menos 150$, pero menos de $180, 60 ganan menos de 210$, 110 ganan menos de $240, 180 ganan menos de $270 y el 10% restante de empleados gana a lo más $ 300; reconstruir la distribución y graficar su polígono de frecuencias.

Rp. X_{min}=150, A=30, frec. absolutas:20, 40, 50, 70, 20.

16. La demanda diaria de azúcar (en decenas de kilos) recopilada durante ciento noventa días en un supermercado, se tabuló en una distribución de frecuencias simétrica de cinco intervalos de amplitud iguales a 4. Si la marca de clase del intervalo central es igual a 12 y si la curva de frecuencias absolutas satisface la relación:

$$f(x) = -(x-12)^2 + 70$$

reconstruir la distribución y graficar su histograma.

Rp. X_{min}=2, frec. absolutas: 6, 54, 70, 54, 6.

17. Los tiempos de vida útil (en días) de un tipo de batería, se tabuló en una distribución de frecuencias de 5 intervalos de igual amplitud con frecuencias relativas acumuladas: 0.10, 0.25, 0.55, 0.80, 1.00. Determine la distribución de frecuencias absolutas si la tercera frecuencia absoluta acumulada es 11, si la segunda marca de clase es 6, y si el límite inferior del cuarto intervalo es 12.

Rp. X_{min}=0, A=4, n=20, frecuencias: 2, 3, 6, 5,4.

18. Los ingresos mensuales de una muestra de pequeños comerciantes se tabularon en una distribución de frecuencias simétrica de 5 intervalos de igual amplitud resultando: Ingreso mínimo \$125, marca de clase del cuarto intervalo $m_4 = \$300$. Si el 8% de los ingresos son menores que \$165 y el 70% de los ingresos son menores a \$275, ¿qué porcentaje de ingresos son superiores a \$285?.

Rp. A=50, frecuencias 0.1, 0.2, 0.4, 0.2, 0.1, % de casos superiores a \$285 es =0.26.

19. La organización del tiempo, en minutos, que tardaron 100 obreros para ejecutar cierta tarea, ha dado una tabla de frecuencias de cuatro intervalos de igual amplitud cuyo histograma correspondiente es simétrico. Si el intervalo $I_1 = [6, ?]$, la frecuencia absoluta: $f_2 = 2f_1 + 5$, y si se sabe que el 85% de los obreros demoran menos de 12 minutos. Completar la distribución de frecuencias.

Rp. X_{min}=6, A=2, frec. relativas: 0.15, 0.35, 0.35, 0.15.

20. Los puntajes de una prueba de aptitud se tabularon en una distribución de frecuencias de 6 intervalos de igual amplitud. Si se tienen: marcas de clase, $m_2 = 40$ y $m_4 = 80$, frecuencias: $h_1 = h_6$, $h_3 = h_5$, $h_4 = 0.25$, $h_2 = h_4 - h_1$, $h_3 = h_1 + 0.10$, y $F_6 = 60$, completar la distribución de frecuencias absolutas y graficar el polígono.

Rp. X_{min}=10, A=20, frecuencias: 6, 9, 12, 15, 12, 6.

21. Las notas de un examen se tabularon en una distribución de frecuencias relativas de 3 intervalos de amplitud iguales a 5. Si la nota mínima es igual a 5, el 48% de las notas son menores que 12, y si el 80% de las notas son inferiores a 16, reconstruir la distribución de frecuencias.

Rp. frecuencias 0.30, 0.45, y 0.25

22. El tiempo (en horas) de 120 familias que utilizan su computadora se tabularon en una distribución de frecuencias de 5 intervalos de amplitud iguales a 4, siendo; el tiempo mínimo de uso 2 horas, la primera y segunda frecuencias iguales al 10% y 15% del total de casos respectivamente. Si el 73.75% de las familias lo usaron menos de 17 horas y el 85% menos de 19 horas, determine las frecuencias. Rp. 0.10, 0.15, 0.30, 0.25, 0.20 o 12, 18, 36, 30, 24

23. Los salarios que ofrece una empresa a los practicantes varían entre $150 y 270$. Si los salarios se agrupan en cuatro intervalos de clase de longitudes iguales de manera que el 40% de los practicantes tienen salarios menores o iguales que $195, el 80% tienen salarios menores o iguales que $225 y el 15% tiene salarios mayores que $232.50.

a) Hallar el porcentaje de practicantes en cada intervalo.

b) Si el ingreso mínimo se fija en $240 y la empresa aumenta una misma cantidad a todos los practicantes de modo que el 20% supere el ingreso mínimo. ¿cuánto sería el aumento?.

Rp. a) frecuencias: 0.10, 0.60, 0.20, 0.10, b) $15.

24. El consumo mensual de agua de 150 hogares, se tabularon en una distribución de frecuencias *simétrica* de 6 intervalos, siendo las frecuencias: $f_2 = 25$, $F_3 = 75$, $F_5 = 130$. Si el límite inferior del sexto intervalo igual a 60, y si el 75% de los consumos son mayores de 43.5m^3, completar la distribución de frecuencias. Rp. X_{min}=35, A=5, frecuencias: 20, 25, 30, 30, 25, 20.

Otras Gráficas

25. La siguiente tabla muestra la superficie (en millones de millas cuadradas) de los océanos.

Océano :	Pacífico	Atlántico	Indico	Antártico	Artico
Superficie:	70	41	28	7	4

Identificar la variable, y represente los datos mediante dos gráficos diferentes.

Rp.gráfica: barras y circular

26. Construya una gráfica adecuada que permita comparar la predilección de los estudiantes por las carreras de ciencias en tres universidades si se tienen los siguientes datos.

Universidades	Alumnos que prefieren ciencias	Total de alumnos
A	300	6000
B	200	4000
C	180	7200

Rp. graficar las frec. Relativas 300/6000=0.05, 200/4000=0.05, 180/7200=0.025

27. Se ha clasificado a un grupo de personas de acuerdo a su ocupación y procedencia. La distribución resultó la siguiente.

	Costa	Sierra	Selva
Agricultores	15	16	7
Mineros	5	9	4
Técnicos	13	8	2
Obreros	16	11	4

a) Haga un gráfico para representar la distribución de las personas por su ocupación.

b) Haga un gráfico para comparar la región de procedencia de las personas según su ocupación.

Rp. grafica a) barras totales, b) barras agrupadas.

28. El volumen de exportación de cobre, en miles de toneladas, durante el período 95-99 se dan en la tabla que sigue. Trazar un gráfico para

Año	Gran minería	Mediana minería	Pequeña minería
1995	30	30	30
1996	50	50	30
1997	80	60	43
1998	60	40	42
1999	50	45	40

a) Mostrar la evolución de las exportaciones.

b) Ver el tipo de minería que determina principalmente la tendencia de las exportaciones.

c) Mostrar la proporción de cada tipo de minería respecto al total de exportaciones por año.

Rp. c) barras por partes componentes.

DIAGRAMA DE TALLO (o TRONCO) Y HOJAS

Es una técnica que se usa para organizar datos sin perder la identidad de cada dato observado, como sí ocurre en una distribución de frecuencias por intervalos.

El diagrama de tallo y hojas se construye partiendo cada dato numérico en dos. El tallo que consiste del dígito o los dígitos iniciales y las hojas que consisten de los dígitos restantes del dato. Usualmente se eligen entre 5 y 20 tallos

Los tallos ordenados son ubicados en forma vertical. Las hojas ordenadas de cada tallo son ubicadas horizontalmente.

EJEMPLO

Con los siguientes datos (ingresos quincenales en dólares del ejemplo 1.3):

63	89	36	49	56	64	59	35	78
43	53	70	57	62	43	68	62	26
64	72	52	51	62	60	71	61	55
59	60	67	57	67	61	67	51	81
53	64	76	44	73	56	62	63	60

a) Desarrolle un diagrama de tallo y hojas.
b) Halle el porcentaje de ingresos quincenales inferiores a $52
c) ¿Cuál es el valor de en medio o central?
d) ¿Cuántos valores están entre 50 y 65?

SOLUCION.

a) Utilicemos el primer dígito de cada dato como tallo y el segundo como hoja. Para el número 63 por ejemplo, el tallo es 6 y la hoja es 3. Como el dato mínimo es 26 y el dato máximo es 89, entonces los tallos empiezan en 2 y terminan en 8. Después de organizar todos los datos, el diagrama de tallo y hojas resulta:

Tallo	Hojas
2	6
3	56
4	3349
5	112335667799
6	000112222334447778
7	012368
8	19

b) El porcentaje de ingresos quincenales inferiores a $52 es (9/45)=0.2 o 20%.
c) El valor de en medio o central es: $61
d) Hay 26 valores entre 50 y 65

EJEMPLO

Los siguientes datos representan el periodo de duración en meses de 32 baterías "DURA" doble A:

3.3	4.0	6.0	4.2	6.0	5.4	4.5
1.5	7.0	6.5	7.4	5.2	5.7	6.2
5.5	5.2	6.8	3.8	2.4	3.6	
5.0	6.2	5.3	6.5	5.5	6.0	
2.8	7.1	6.7	4.7	5.6	5.9	

a) Desarrolle un diagrama de tallo y hojas.
b) ¿Cuál es el valor de en medio?
c) ¿Cuántas baterías duraron entre 2.9 y 5.8 meses?

SOLUCION.

a) Utilicemos dígito entero de cada dato como tallo y el dígito decimal como hoja. Para el número 5.2 por ejemplo, el tallo es 5 y la hoja es 2. Como el dato mínimo es 1.5 y el dato máximo es 7.4, entonces los tallos empiezan en 1 y terminan en 7.
Después de organizar todos datos, el diagrama de tallo y hojas resulta:

Tallo	Hojas
1	5
2	48
3	368
4	0257
5	0223455679
6	000225578
7	014

b) El valor de en medio es: 5.5 meses
c) Hay 16 baterías que duraron entre 2.9 y 5.8 meses.

EJEMPLO

El siguiente es el diagrama de tallo y hojas de los valores (un entero y dos decimales) de una variable continua. El tronco consiste de un entero y un decimal:

Tallo	Hojas
35	245
36	3567
37	34556889
38	0024455
39	358

a) ¿Cuántos datos se observaron?. Indique el mínimo y el máximo.
b) ¿Cuál es el valor del centro?

Rp. a) 25 datos, Min = 3.53, Max = 3.98, b) 3.78

Capítulo 2

MEDIDAS DE POSICION

2.1 Introducción

Los datos organizados en una distribución de frecuencias destacan sus características más esenciales, como marcas de clases, centro, forma de distribución (asimétrica, simétrica,) etc.. Sin embargo, los indicadores que describen a los datos en forma más precisa, deben calcularse. Estos indicadores resumen los datos en medidas descriptivas que se refieren a la **centralización** o posición, a la **dispersión** o variación, a la **asimetría**, y a la **curtosis** de los datos.

Las medidas de posición reflejan la tendencia central y la localización de los datos.

Las medidas de tendencia central, denominados también **promedios**, ubican el centro de los datos, como la *media aritmética*, la *media geométrica*, la *media armónica* y la *mediana*.

Las medidas de localización indican el lugar de los datos más frecuentes (*moda*) o de los menos frecuentes a partir de los *cuantiles*.

El lector debería correr paquetes de computo entre otros el *MCEST* para las aplicaciones de este capitulo.

2.2 Mediana

Definición. La *mediana* o valor mediano de una serie de valores observados es el número M_e que separa a la serie de datos *ordenados* en forma creciente (o decreciente) en dos partes de igual número de datos.

La mediana es la medida promedio que depende del número de datos ordenados y no de los valores de estos datos.

2.2.1 Cálculo de la mediana

1) Mediana de datos no tabulados

Se sigue el método de percentiles de la página 26. Esto es, para determinar la mediana de n valores no tabulados de una variable cuantitativa X :

1) Se ordenan los datos en forma creciente.
2) Luego, se ubica el valor central Me. Si n es impar, la mediana es el dato ordenado del centro. Pero si n es par, la mediana es la semisuma de los dos valores ordenados centrales.

EJEMPLO 2.1
Calcular la mediana para las siguientes series de datos.

a) 120, 3, 14, 1, 99, 7, 30, 2,000, 16

b) 30, 77, 3, 300, 36, 11, 10,000, 29

SOLUCION

a) La serie ordenada de los 9 datos es :

1, 3, 7, 14, 16, 30, 99, 120, 2,000.

La mediana es el quinto dato ordenado que divide a la serie en 2 grupos de 4 datos cada uno. Esto es, $Me = 16$.

b) La serie ordenada de los 8 datos es:

3, 11, 29, 30, 36, 77, 300, 10,000.

la mediana en este caso, puede ser cualquier número situado entre 30 y 36, ya que este dividirá a los datos en dos grupos de 4 datos cada uno. Pero, para evitar la infinidad de valores, se elige como mediana la semisuma de los dos valores centrales. Esto es, $Me = (30+36)/2 = 33$.

Observar pues, que la mediana no depende de la magnitud de los datos. Depende sólo del número de ellos.

NOTA. El lector puede verificar que la mediana de los 45 ingresos quincenales sin tabular por intervalos del ejemplo 1.3, es igual a 61 dólares.

2) Mediana de datos tabulados

2a) Si los valores de una ***variable discreta*** se tabulan en una distribución de frecuencias de la forma "dato ↔ frecuencia", el cálculo de la mediana se hace siguiendo el procedimiento anterior 1) debido a que los datos están ordenados.

Por **ejemplo**, la mediana para la distribución del número de hijos por familia del ejemplo 1.2 es igual a 2.

2b) Si los valores de la variable se tabulan en una **distribución de frecuencias por intervalos**, la mediana se determina aproximadamente por interpolación a partir de la distribución de frecuencias acumuladas, método ya propuesto en aplicaciones de la ojiva.(capitulo1)

Para calcular la mediana:

Primero se determina el intervalo $I_i = [L_i, U_i[$ que contiene a la mediana Me. Para esto, se determinan las frecuencias acumulada F_i y F_{i-1} de manera que: (ver figura 2.1)

$$F_{i-1} \leq n/2 < F_i.$$

A partir de $n/2$ en el eje vertical, se traza una paralela a la ojiva, y de la ojiva se traza una vertical al eje de los intervalos.

La mediana $Me \in [L_i, U_i[$, intervalo de amplitud A, cuya frecuencia absoluta acumulada es F_i, y la no acumulada es $f_i = F_i - F_{i-1}$.

Segundo, la mediana $Me \in [L_i, U_i[$. Luego,

$$Me = L_i + a$$

donde a se obtiene por interpolación (semejanzas de triángulos ABE y ACD de la figura 2.1) comparando intervalos con frecuencias:

$$\frac{a}{A} = \frac{\frac{n}{2} - F_{i-1}}{f_i}, \quad \text{de donde} \quad a = \frac{\frac{n}{2} - F_{i-1}}{f_i} A,$$

Luego,
$$Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} A$$

Donde:

L_i es el limite inferior del intervalo de la mediana.

n es el número de datos observados.

F_{i-1} es la frecuencia acumulada absoluta del intervalo inmediatamente anterior al intervalo de la mediana.

f_i es la frecuencia absoluta del intervalo de la mediana.

A es la amplitud del intervalo de la mediana.

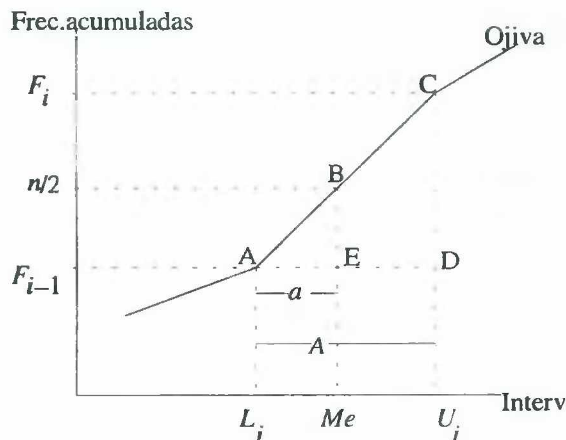


Fig. 2.1 Determinación de la mediana por interpolación

EJEMPLO 2.2

Calcular la mediana para la muestra de los 45 ingresos quincenales tabulados del ejemplo 1.3

SOLUCION

Los datos tabulados del ejemplo 1.3 se repiten en el cuadro 2.1, donde $n = 45$. $n/2=22.5$.

Cuadro 2.1. Cálculo de la mediana

Ingresos I_i	Número de personas f_i	Frec. acumuladas F_i
[26,34[	1	1
[34,42[	2	3
[42,50[	4	7
[50,58[	10	17
[58,66[	16	33
[66,74[	8	41
[74,82[	3	44
[82,90]	1	45
Total	45	

El valor 22.5 está entre las frecuencias acumulada 17 y 33, luego, la mediana, $Me \in [58, 66[$.

Además $L_i = 58$, $F_{i-1} = 17$, $f_i = 16$, $A = 8$,

Luego,

$$Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} A = 58 + \left(\frac{22.5 - 17}{16}\right) \times 8 = 58 + 2.75 = 60.75.$$

NOTA (Cálculo de la mediana para frecuencias relativas)

Si en lugar de las frecuencias absolutas se utilizan las frecuencias relativas (o porcentajes), entonces, haciendo $h_i = f_i / n$, $H_{i-1} = F_{i-1} / n$ en la fórmula de la mediana, se tiene:

$$Me = L_i + \left(\frac{1/2 - H_{i-1}}{h_i}\right) A = 58 + \left(\frac{1/2 - 17/45}{16/45}\right) \times 8 = 60.75$$

2.2.3 Propiedades de la mediana

1) La mediana, sólo depende del número de datos ordenados y no del valor de los datos. Por lo tanto, no es sesgada por algún valor grande o pequeño.

2) La mediana puede ser calculada para distribuciones de frecuencia con intervalos de diferente amplitud, siempre que se pueda determinar el límite inferior del intervalo de la mediana, L_i.

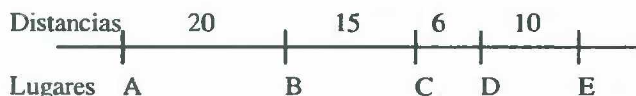
3) La mediana puede ser calculada para variables con valores en escala ordinal.

4) La suma de las diferencias (en valor absoluto) de *n* datos con respecto a su mediana es mínima. En el caso de datos sin tabular,

$$\sum_{i=1}^{n} |x_i - c| = \text{mínimo, si } c \text{ es la } \textit{mediana} \text{ de los } x_i.$$

EJEMPLO 2.3

Cinco personas que viven en lugares situados a distancias en kilómetros a lo largo de una carretera en línea recta como se indica en la figura que sigue, deben reunirse en algún punto de la carretera. Determine el lugar de reunión de manera que el costo total del transporte sea mínimo, si el costo de cada transporte es proporcional al recorrido.



SOLUCION

Sea A el origen, entonces, las coordenadas de A, B, C, D y E son respectivamente: $x_1 = 0$, $x_2 = 20$, $x_3 = 35$, $x_4 = 41$ y $x_5 = 51$.

Sea K el lugar de reunión. Dado que el costo es proporcional al recorrido, el *costo total* del transporte es:

$$\sum_{i=1}^{5} |x_i - K|.$$

Este costo total es mínimo, si K es la mediana de los 5 valores 0, 20, 35, 41, 51, esto es si $K = 35$. Luego, deben reunirse en el lugar C a 35 kilómetros de A.

2.3 Moda

Definición. La *moda* de una serie de datos es el valor Mo, **que se define como el dato que más veces se repite.**

La moda no siempre existe y si existe, no siempre es única.

En matemática, la moda es el valor de la variable en el que existe un máximo absoluto (o dos o más máximos relativos iguales).

La moda es una medida promedio que se usa cuando se quiere señalar el valor más común de una serie de datos. Por ejemplo, los comerciantes se estoquean con productos que están de moda.

La moda es el promedio menos importante debido a su ambigüedad.

EJEMPLO 2.4

La moda de los datos:

a) 7, 9, 7, 8, 7, 4, 7, 13, 7 es igual a 7. Esta serie de datos es *unimodal*

b) 5, 3, 4, 5, 7, 3, 5, 6, 3 es igual tanto a 3, como a 5. Esta serie de datos es *bimodal*.

c) 31, 11, 12, 19 no existe. (También vale decir que cada uno de los datos es una moda).

NOTA. La moda de los 45 ingresos quincenales sin tabular del ejemplo 1.3, es igual a 62 dólares (verificar!).

Moda de datos tabulados por intervalos

Para calcular la moda de n datos tabulados por intervalos, primero se determina el intervalo que contiene a la moda, esto es, el intervalo que tiene la mayor frecuencia (*intervalo modal*). Luego se utiliza la fórmula

$$Mo = L_i + \left(\frac{d_1}{d_1 + d_2} \right) A \ ,$$

donde:

L_i es el límite inferior del intervalo modal.

$d_1 = f_i - f_{i-1}$, esto es, d_1 es igual a la frecuencia del intervalo modal menos la frecuencia del intervalo inmediatamente anterior.

$d_2 = f_i - f_{i+1}$, esto es, d_2 es igual a la frecuencia del intervalo modal menos la frecuencia del intervalo inmediatamente posterior.

A es la amplitud del intervalo modal.

NOTA. La fórmula de la moda sólo se aplica en distribuciones con una sola frecuencia máxima.

EJEMPLO 2.5

Calcular la moda de los 45 ingresos quincenales del ejemplo 1.3 tabulados por intervalos.

SOLUCION.

Los datos tabulados por intervalos del ejemplo 1.3 se repiten en el cuadro 2.1. Aquí se observa que $Mo \in [58,66[$. Además,

$$L_i = 58, \quad d_1 = 16 - 10 = 6, \ d_2 = 16 - 8 = 8, \quad A = 8.$$

Luego, la moda de la distribución es:

$$Mo = L_i + \left(\frac{d_1}{d_1 + d_2} \right) A = 58 + \left(\frac{6}{6+8} \right) \times 8 = 61.43.$$

NOTA: (Cálculo de la moda con frecuencias relativas)

Si en lugar de las frecuencias absolutas se utilizan las frecuencias relativas, se tiene:

$$Mo = L_i + \left(\frac{d_1}{d_1 + d_2} \right) A = 58 + \left(\frac{6/45}{(6+8)/45} \right) \times 8 = 61.43.$$

2.4 Media aritmética

Definición. La *media aritmética*, denominada simplemente media, es la suma de los valores observados de la variable, dividido por el número de observaciones.

Para valores de una variable X observados en una *muestra*, la media aritmética se denota por $\overline{x}$.

2.4.1 Cálculo de la media aritmética

1) Media aritmética de datos no tabulados

La media de n valores $x_1, x_2, ..., x_n$, de la variable cuantitativa X, observados en una muestra es el número:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

EJEMPLO 2.6

Utilizando los datos sin tabular del ejemplo 1.3 calcular la media aritmética.

SOLUCION

La suma de los 45 ingresos sin tabular es \$2682. Luego,

$$\overline{x} = \frac{\sum_{i=1}^{45} x_i}{45} = \frac{\$2682}{45} = \$59.6$$

2) Media aritmética de datos tabulados

2a) Media para datos tabulados de variable discreta

Si n valores de una variable estadística discreta X se clasifican en k valores distintos $x_1, x_2, ..., x_k$ con frecuencias absolutas respectivas $f_1, f_2, ..., f_k$, entonces, su media aritmética es el número:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}.$$

EJEMPLO 2.7

Calcular la media aritmética de la distribución del número de hijos por familia del ejemplo 1.2.

SOLUCION

La distribución de frecuencias del ejemplo 1.2 se repite en el cuadro 2.2,donde se ha incluido una columna de productos $f_i x_i$

Cuadro 2-2. Cálculo de la media del número de hijos por familia

Valores de X	frecuencias	Productos
x_i	f_i	$f_i x_i$
0	1	0
1	4	4
2	7	14
3	6	18
4	2	8
Total	20	44

resultando: $$\overline{x} = \frac{\sum_{i=1}^{5} f_i x_i}{20} = \frac{44}{20} = 2.2.$$

2b) Media para datos tabulados por intervalos

Si n valores de alguna variable X están tabulados en una distribución de frecuencias de k intervalos, donde:

$m_1, m_2, \ldots, m_k$ son las marcas de clase , y

$f_1, f_2, \ldots, f_k$ son las frecuencias absolutas respectivas,

entonces, su media aritmética es el número:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n}.$$

EJEMPLO 2.8

Calcular la media de la distribución de frecuencias de los 45 ingresos quincenales del ejemplo 1.3.

SOLUCION

La distribución de frecuencias del ejemplo 1.3 se repite en el cuadro 2.3. donde se ha incluido una columna de productos $f_i m_i$

Resultando, $$\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n} = \frac{\$2702}{45} = \$60.04.$$

Cuadro 2.3. Cálculo de la media aritmética de los ingresos de una muestra de 45 personas

ingresos I_i	Marcas m_i	N#.personas f_i	Productos $f_i m_i$
[26,34[	30	1	30
[34,42[	38	2	76
[42,50[	46	4	184
[50,58[	54	10	540
[58,66[	62	16	992
[66,74[	70	8	560
[74,82[	78	3	234
[82,90]	86	1	86
Total		45	2702

NOTA. Observar que la media de los 45 datos sin tabular es $59.6.La media obtenida por el método de distribución de frecuencias por intervalos es una media aproximada.

NOTA. La media aritmética de datos tabulados, se calcula también, utilizando las frecuencias relativas (o porcentajes), haciendo, $h_i = f_i / n$, $i = 1,2,\ldots,k$. En el caso de datos agrupados por intervalos, se tiene,

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n} = \sum_{i=1}^{k} h_i m_i$$

NOTA (Media aritmética de la población). La media aritmética de una población denotamos por μ. Si la población es finita de tamaño N con valores $x_1, x_2, \ldots, x_N$, la media es el número:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

2.4.2 Propiedades de la media aritmética

1) La suma total de n valores cuya media es $\bar{x}$ es igual a $n\bar{x}$. Para n datos no tabulados y tabulados respectivamente, se tiene:

$$\sum_{i=1}^{n} x_i = n\bar{x}, \qquad \sum_{i=1}^{k} f_i x_i = n\bar{x}.$$

2) Si cada uno de los n valores x_i es transformado en: $y_i = ax_i + b$, siendo a y b constantes, entonces, la media de los n valores y_i es (verificar!):

$$\bar{y} = a\bar{x} + b.$$

Como casos particulares se tiene:

Si $y_i = b$, entonces, $\bar{y} = b$. Esto es, si los n datos son iguales a una constante, entonces su media es igual a esa constante.

Si $y_i = x_i + a$, entonces, $\bar{y} = \bar{x} + b$. Esto es, si a cada dato se suma una constante la media queda sumada por esa constante.

Si $y_i = ax_i$, entonces, $\bar{y} = a\bar{x}$. Esto es, si a cada dato se multiplica por una constante, la media queda multiplicada por esa constante.

3) La suma algebraica de las desviaciones de n datos x_i con respecto a su media $\bar{x}$, es igual a cero. Para datos no tabulados, y tabulados, se tiene respectivamente:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0, \qquad \sum_{i=1}^{k} f_i (x_i - \bar{x}) = 0.$$

4) La suma de los cuadrados de las desviaciones de n datos con respecto a su media es mínima. Para datos no tabulados, por ejemplo,

$$\sum_{i=1}^{n} (x_i - c)^2 = \text{mínima}, \quad \text{si } c = \bar{x}.$$

En efecto, si c es cualquier número real,

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - c)\sum_{i=1}^{n}(x_i - \bar{x}) + n(\bar{x} - c)^2.$$

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2, \quad \text{ya que } \sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

$$\sum_{i=1}^{n}(x_i - c)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \text{ya que} \quad n(\bar{x} - c)^2 \geq 0.$$

NOTA. (Media aritmética ponderada).

La media aritmética de los valores $x_1, x_2, ..., x_k$ ponderada por los pesos $w_1, w_2, ..., w_k$ es el número:

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + ... + w_k x_k}{w_1 + w_2 + ... + w_k} = \sum_{i=1}^{k} w_i x_i \Big/ \sum_{i=1}^{k} w_i \,.$$

Por **ejemplo**, si un alumno en el semestre anterior ha obtenido 11 en el curso A de 5 créditos, 13 en el curso B de 4 créditos, y 16 en el curso C de 3 créditos, entonces, su promedio (ponderado por los créditos) es,

$$\bar{x} = \frac{11 \times 5 + 13 \times 4 + 16 \times 3}{5 + 4 + 3} = \frac{155}{12} = 12.92.$$

En realidad, toda media aritmética es ponderada. En el caso de datos no tabulados, el peso de cada valor de la variable es igual a uno.

Los pesos pueden ser también números relativos o porcentajes.

Por **ejemplo**, si en este mes el aumento de los alimentos fue del 5%, de vivienda el 10% y de educación 8 %. Entonces, **el aumento promedio** en los tres rubros para una persona que gasta el 40% de su sueldo en alimentos, el 35% en vivienda y el 25% en estudios está dado por:

$$\bar{x} = 0.05 \times 0.4 + 0.10 \times 0.35 + 0.08 \times 0.25 = 0.075.$$

Pero el aumento promedio en los tres rubros para una persona que gasta S/.1,200 en alimentos, S/.600 en vivienda y S/. 1,000 en estudios está dado por:

$$\bar{x} = \frac{0.05 \times 1200 + 0.10 \times 600 + 0.08 \times 1000}{1200 + 600 + 1000} = \frac{200}{2800} = 0.0714.$$

NOTA. (Media aritmética total a partir de medias de sus partes).

La media del total de datos es igual a la media ponderada de sus partes donde los pesos son los tamaños respectivos.

Esto es, si *n* datos de variable *X*, consiste de *k* partes de tamaños respectivos $n_1, n_2, ..., n_k$ y medias respectivas $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$, entonces, la media del total de datos es igual a:

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + ... + n_k\bar{x}_k}{n_1 + n_2 + ... + n_k}.$$

Por **ejemplo**, si en un examen, 110 alumnos de la sección H1 obtuvieron una media de 12.6 y 120 alumnos de la sección H2 obtuvieron un promedio de 13.48, entonces, los 230 alumnos en conjunto obtuvieron el promedio

$$\bar{x} = \frac{110 \times 12.6 + 120 \times 13.48}{110 + 120} = 13.06.$$

EJEMPLO 2.9

Los sueldos del mes de enero de 200 empleados de una empresa tienen una media de \$230.

a) Si el 60% de los empleados son hombres (el resto son mujeres) y tienen un sueldo promedio de \$ 250, ¿cuánto es el sueldo medio de las mujeres en enero?

b) Si en el mes de julio, se propone un aumento del 30% a cada sueldo de enero más una bonificación de \$30 ¿cuánto dinero adicional necesitará la empresa para pagar los sueldos de julio?

SOLUCION

a) Sean $\bar{x}$, $\bar{x}_1$ y $\bar{x}_2$ las medias total, de hombres y de mujeres respectivamente. Entonces,

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = h_1\bar{x}_1 + h_2\bar{x}_2, \text{ donde } h_1 = \frac{n_1}{n_1 + n_2}, \quad h_2 = \frac{n_2}{n_1 + n_2}$$

Luego, $230 = 0.6 \times 250 + 0.4 \times \bar{x}_2$, entonces, $\bar{x}_2 = 200$.

b) Sean X: sueldos de enero, Y: sueldos de julio. Entonces,

$$Y = X + 0.30X + 30 = 1.3X + 30$$

$$\bar{y} = 1.3\bar{x} + 30 = 1.3(230) + 30 = 329$$

Total de dinero para pagar sueldos de enero = $n(\bar{x})$ = 200(230) = \$46,000

Total de dinero para pagar sueldos de julio = $n(\bar{y})$ = 200(329) = \$65,800

Dinero adicional para pagar sueldos de julio = \$65,800 – \$46,000 =\$19,800.

EJEMPLO 2.10

Con los datos del ejemplo 2.3. Determine el lugar de reunión de manera que el costo total del transporte sea mínimo, si el costo de cada transporte es proporcional al cuadrado del recorrido.

SOLUCION

Sea K el lugar de reunión. Dado que el costo del transporte es proporcional al cuadrado del recorrido, el *costo total* del transporte es:

$$\sum_{i=1}^{5}(x_i - K)^2.$$

Este costo total será mínimo, si K es la media de los 5 valores 0, 20, 35, 41, 51, esto es si $K = 29.4$. Luego, deben reunirse en un lugar a 29.4 Km. de A ubicada en la carretera.

EJEMPLO 2.11

Un conjunto de n artículos cuyos valores de venta serían de \$5 en el 30% de los casos, \$7 en el 45 % de los casos y 10\$ en el 25% de los casos, tienen un costo de producción fijo de \$$k$. Hallar el valor de k si se quiere hacer una inversión mínima y si se supone que la inversión es igual a la suma de los cuadrados de todas las utilidades.

SOLUCION

La inversión es: $(5-k)^2$ en el 30% de los casos + $(7-k)^2$ en el 45% de los casos + $(10-k)^2$ en el 25% de los casos. La inversión es mínima, si k es la media de los datos: \$5 con frecuencia 0.30, \$7 con frecuencia 0.45 y \$10 con frecuencia 0.25. Esto es, si

$$k = 5 \times 0.3 + 7 \times 0.45 + 10 \times 0.25 = \$7.15$$

NOTA. (Defectos de la media aritmética)

1. La media aritmética depende de todos los valores observados, en consecuencia es **"afectada"** o **"sesgada"** por valores extremos. Por **ejemplo**, la media aritmética de los grupos:

a) 55, 56, 57, 58, 59, 60 es igual a $\bar{x}_1 = 345/6 = 57.5$.

b) 55, 56, 57, 58, 59, 100 es igual a $\bar{x}_2 = 385/6 = 64.2$.

c) 55, 56, 57, 58, 59, 0 es igual a $\bar{x}_3 = 285/6 = 47.5$.

Como se puede observar, la media aritmética es sesgada por los valores extremos: 100 en el grupo b) y 0 en el grupo en c).

2. La media aritmética puede ser calculada también en distribución de frecuencias por intervalos de amplitud diferentes, ***siempre que puedan determinarse los***

puntos medios (marcas) de las intervalos. Por **ejemplo**, no se puede calcular la media a partir de una distribución de frecuencias como del cuadro 2.4. (También no se puede calcular la mediana y la moda, ¿por qué?)

Cuadro 2.4

Intervalos	Frecuencias
menor o igual que 20	50
[20, 25[	20
25 o más	10

2.5 Relación entre media, mediana y moda

1. Si la distribución de frecuencias es simétrica, entonces, la media, la mediana y la moda tienen el mismo valor (figura 2.2(a)). Esto es,

$$\overline{x} = Me = Mo.$$

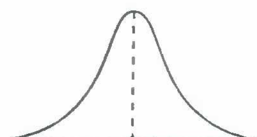


fig. 2.2. (a)

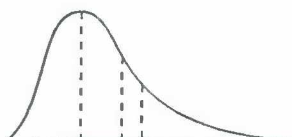


fig. 2.2. (b)

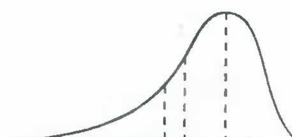


fig. 2.2. (c)

2. Si la distribución es asimétrica de cola a la derecha, entonces, la moda es menor que la mediana y esta a su vez es menor que la media (figura 2.2(b)). Es decir :

$$Mo < Me < \overline{x}.$$

3. Si la distribución es asimétrica de cola a la izquierda, entonces, la relación es (figura 2.2(c)):

$$\overline{x} < Me < Mo.$$

4. Para distribuciones unimodales y asimétricas, se tiene la siguiente relación empírica:

$$\overline{X} - Mo \cong 3(\overline{X} - Me).$$

5. Los tres promedios pueden calcularse también para distribuciones de frecuencias con intervalos de diferente longitud, siempre que puedan determinarse o las marcas de clase (para la media) o el límite inferior L_i del intervalo (para la mediana y la moda).

2.6 Uso de los promedios

1. De los promedios definidos, la media aritmética se usa con más frecuencia por su mejor tratamiento algebraico. Pero no siempre es un buen promedio.
2. Si la distribución de frecuencias es simétrica (o "casi" simétrica), la media, o la mediana o la moda es el promedio representativo, pues, en este caso, los tres promedios son iguales (o casi iguales).
3. Si la distribución tiene marcada asimetría, entonces, la mediana es la medida promedio más representativa.

2.7 Otras medias

La media geométrica

Definición. La media geométrica de n valores positivos $x_1, x_2, ..., x_n$ es el número $\bar{x}_G$ que se define como la raíz enésima del producto de estos n valores. Esto es,

$$\bar{x}_G = \sqrt[n]{x_1 \times x_2 \times ... \times x_n}.$$

Por ejemplo, la media geométrica de los valores 3, 9, 27 es igual a:

$$\bar{x}_G = \sqrt[3]{3 \times 9 \times 27} = 9.$$

La media geométrica se utiliza para promediar: razones (a/b), índices (a/b en %), proporciones ($a/(a+b)$), tasas de cambio $(a-b)/b$, que varían con el tiempo, entre otros.

EJEMPLO 2.12

Si una producción ha experimentado un crecimiento del 30% del primero al segundo año y un incremento del 35% del segundo al tercer año y un decrecimiento del 15% del tercer al cuarto año.

a) Calcular la tasa promedio de crecimiento de los 3 últimos años.

b) Calcular la producción del quinto año, si la del primer año es 100.

SOLUCION.

a) Tomando como producción base 100 para el primer año.

En el segundo año, el porcentaje de crecimiento es de 30%, la producción es: $100 + 0.30 \times 100 = 130$ y la tasa de crecimiento es $130/100 = 1.30$.

En el tercer año, el porcentaje de crecimiento es de 35%, la producción es: $130 + 0.35 \times 130 = 175.5$ y la tasa de crecimiento es $175.5/130 = 1.35$.

En el cuarto año, el porcentaje de crecimiento es de -15%, la producción es: $175.5 - 0.15(175.5) = 149.175$ y la tasa de crecimiento es $149.175/175.5 = 0.85$.

Cuadro 2.5. Cálculo de la media geométrica

Año	% Crecimiento	Producción	Tasas
1	–	100	–
2	30%	100+0.3(100)=130	130/100=1.30
3	25%	130+0.35(130)=175.5	175.5/130=1.35
4	–15%	175.5–0.15(175.5)=149.175	149.175/175.5=0.85

El promedio de las tasas de aumento durante los tres últimos años es la media geométrica:

$$\bar{x}_G = \sqrt[3]{1.30 \times 1.35 \times 0.85} = 1.1426 = 1 + \frac{14.26}{100}.$$

Esto es, el porcentaje promedio de crecimiento es de 14.26%.

b) La producción para el quinto año es igual a:

$$149.175 + 0.1426 \times 149.175 = 170.44$$

EJEMPLO 2.13.

Suponga que la población de una ciudad aumentó de 10,000 a 12,600 en el período de 1980 a 1984 como se indica en el cuadro 2.5. Determine la tasa media del crecimiento.

Cuadro 2.5. Cálculo de la media geométrica

Año	Población	Tasas de cambio(X) (año base 1980)	log(X)
1980	10,000	–	–
1981	10,500	1.050	0.0212
1982	11,200	1.067	0.0282
1983	12,000	1.071	0.0298
1984	12,600	1.050	0.0212
total			0.1004

SOLUCION.

Para simplificar los cálculos, se puede utilizar logaritmos en base 10,

$$\log(\bar{x}_G) = \frac{\log(x_1) + \log(x_2) + ... + \log(x_n)}{n} = \frac{\sum_{i=1}^{n} \log(x_i)}{n}.$$

de donde resulta:

$$\bar{x}_G = anti\log\left(\sum_{i=1}^{n} \log(x_i) \Big/ n\right).$$

Para calcular la tasa media de cambio por año escogemos como base el año 1980. Las tasas de cambio se obtienen dividiendo cada dato entre el dato del año anterior. Aplicando la media geométrica a las tasas se tiene:

$$\log(\bar{x}_G) = \frac{\sum_{i=1}^{n} \log(x_i)}{n} = \frac{0.1004}{4} = 0.0251, \quad \bar{x}_G = anti\log(0.0251) = 1.0595.$$

Luego, la tasa promedio de crecimiento es de 5.95% por año.

La media armónica

Definición. La media armónica de n valores no nulos $x_1, x_2, ..., x_n$ es el número , $\bar{x}_H$ que se define como el recíproco de la media aritmética de los recíprocos de esos n valores.
Esto es,

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Por ejemplo, la media armónica de los valores 6, 4, y 3 es igual a 4.

La media armónica se utiliza para obtener promedios de valores que están en relación inversa como la velocidad y el tiempo. En general, se usa para obtener el promedio de un conjunto de valores expresados en forma de tasas de unidades de un tipo por unidades de otro tipo (por ejemplo Km/h).

NOTA. La media armónica es siempre menor que la media geométrica. Esta a su vez es menor que la media aritmética. Esto es,

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

EJEMPLO 2.14

Una persona manejando su automóvil recorre los primeros 10 Km. a 60 Km. por hora y los siguientes 10 Km. a 70 Km. por hora, calcular la velocidad media.

SOLUCION.

Para recorrer los primeros 10 Km. usa 10/60 horas. Para recorrer los siguientes 10 Km. usa 10/70 horas. Por lo tanto, para cubrir los 20 Km. (10 + 10) se emplearon (10/60)+(10/70) horas con un promedio de velocidad de:

$$\bar{x}_H = \frac{\text{Recorrido total}}{\text{Tiempo total}} = \frac{10+10}{\frac{10}{60}+\frac{10}{70}} = \frac{2}{\frac{1}{60}+\frac{1}{70}} = 64.6 \text{ km/h.}$$

EJEMPLO 2.15

Una empresa de transporte gasta \$400 en latas de aceite que cuestan \$10 la docena; \$500 en latas que cuestan \$12.5 la docena; \$600 más en latas que cuestan \$20 la docena y \$300 en otras que cuestan \$25 la docena. Calcular el costo promedio por docena de las latas de aceite.

SOLUCION

$$\bar{x}_H = \frac{\text{Total de costos}}{\text{Total de docenas}} = \frac{400+500+600+300}{\frac{400}{10}+\frac{500}{12.5}+\frac{600}{20}+\frac{300}{25}} = \$14.75$$

EJEMPLO 12.16

Una propiedad de la media es: "si $x_1, x_2, \ldots x_n$ tienen media $\bar{x}$ y si la función $f(x)$ es cóncava hacia abajo, entonces $f(\bar{x}) \geq \frac{1}{n}\sum_{i=1}^{n} f(x_i)$".

Utilizando la propiedad indicada pruebe que $\bar{x}_G \leq \bar{x}$, donde $\bar{x}_G$ es la media geométrica de los datos.

SOLUCION

$\bar{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n}$, entonces,

$\log(\bar{x}_G) = \frac{1}{n}\sum_{i=1}^{n} \log(x_i) \leq \log(\bar{x})$, donde $f = \log$. Además,

$\log(\bar{x}_G) \leq \log(\bar{x})$, implica. $\bar{x}_G \leq \bar{x}$, (Dado que $f = \log$ es inyectiva)

EJERCICIOS

1. Los costos de fabricación, en soles, de diez objetos son los siguientes:

9.35, 9.46, 9.20, 9.80, 9.77, 9.00, 9.99, 9.36, 9.50, 9.60.

si el precio de venta de cada objeto es 3 veces su costo de fabricación menos 5 soles, calcular la utilidad media por objeto.

Rp $y_i=3x_i-5$, $\bar{y} = 3\bar{x} - 5$, $\bar{x} = 9.503$, $\bar{y} = 23.509$, utilidad media=S/.14.006

2. En una evaluación, 5 alumnos tienen cada uno nota 12, y un alumno tiene 18. Si se indica como nota promedio 13, ¿qué nota promedio es?, ¿es el promedio adecuado?, ¿cuánto es el promedio adecuado?.

Rp. media, no, *Me* = 12.

3. De las edades de cuatro personas, se sabe que la media es igual a 24 años, la mediana es 23 y la moda es 22. Encuentre las edades de las cuatro personas.

Rp. 22, 22, 24 y 28.

4. De la curva de frecuencias de los sueldos de 30 empleados de una empresa, se sabe que $Mo = \$200$, $Me = \$220$, y $\overline{x} = \$250$. Califique como verdadera o falsa las siguientes afirmaciones, justificando su respuesta:

a) El sueldo más frecuente es de $200 y más de la mitad de los empleados gana más de esa cantidad.

b) Con una suma de $3,300 se asegura el pago de la mitad de los empleados y con $7,500 el de todos los empleados. Rp. a) si, b) si.

5. Para calcular el suministro de agua que una ciudad requiere mensualmente, se escogen 15 familias de la ciudad, resultando los siguientes consumos en metros cúbicos:

11.2, 21.5, 16.4, 19.7, 14.6,
16.9, 32.2, 18.2, 13.1, 23.8,
18.3, 15.5, 18.8, 22.7, 14.0.

Si en la ciudad hay 5,000 familias, ¿cuántos metros cúbicos de agua se requieren mensualmente si el consumo promedio por familia permanece igual?.

Rp. media=276.9/15=18.46, consumo estimado=5,000(18.46)=92300

6. El sueldo promedio de 200 empleados de una empresa es S/400. Se proponen dos alternativas de aumento:

a) S/. 75 a cada uno, b) 15% de su sueldo más 10 soles a cada uno.

Si la empresa dispone a lo más de S/. 94,000 para pagar sueldos, ¿cuál alternativa es más conveniente?

Rp. total a) $n\overline{x} = 200 \times 475 = 95000$ b) $n\overline{y} = 200 \times 470 = 94000$, b.

7. Al calcular la media de 125 datos, resultó 42. Un chequeo posterior mostró que en lugar del valor 12.4 se introdujo 124. Corregir la media.

Rp. media corregida=41.1072.

8. Las ventas de un distribuidor de automóviles, en cierto periodo, ascendieron a la cantidad de $ 1'650,000, vendiendo 50 automóviles nuevos a un precio promedio de $ 13,000 y algunos carros usados con un precio de $5000 en promedio. ¿Cuál es el promedio de los precios de todos los automóviles que se vendieron?.

Rp. n_2 =200, media=1,650,000/250=$6,600

9. De los horarios de clases de EE.GG.CC. se sabe que ninguno tiene más de 100 o menos de 70 alumnos matriculados. Se sabe que uno de cada 5 tiene 80

alumnos, que el 30% tiene 100 y la mayoría 90 alumnos. Calcular la media aritmética de alumnos por horario.

Rp. $0.2\times80 + 0.5\times90 + 0.3\times100 = 91$.

10. En tres grupos distintos de 100,000, 90,000 y 20,000 personas, el porcentaje de personas con educación superior es 21%, 42% y 40%, respectivamente. Calcular el porcentaje promedio de personas con educación superior

Rp. $(0.21\times100{,}000 + 0.42\times90{,}000 + 0.40\times20000)/210.000 = 0.318$.

11. En un informe (que se supone es correcto) sobre sueldos en todo el país una empresa de estudios de mercados publica la siguiente tabla

	Clase A	Clase B	Clase C	Clase E
% de población	10%	25%	35%	30%
Sueldos	S/. 2500	S/. 1500	S/. 500	S/. 200

y concluye diciendo que "la media de los sueldos en todo el país es S/1175".

a) ¿Qué comentario le merece el informe?. Si no está de cuerdo, ¿cuál sería la corrección?

b) ¿Es la media en este caso el promedio representativo?, si no está de acuerdo, ¿cuánto es el promedio adecuado?.

Rp. a) Media correcta$=2500\times0.1+1500\times0.25+500\times0.35+200\times0.30=860$., b) Mediana=\$500.

12. De una central telefónica salieron 70 llamadas de menos de 3 minutos, promediando 2.3 minutos, 40 llamadas de menos de 10 minutos pero no menos de 3 minutos, promediando 6.4 minutos, y 10 llamadas de al menos 10 minutos, promediando 15 minutos. Calcular la duración promedio de todas las llamadas.

Rp. 4.725 minutos.

13. Cuatro fábricas A, B, C y D, producen un mismo objeto. La fábrica B produce el doble de C, la D 10% menos que la C y la A el 60% menos que la B. Los costos de producción (en dólares) por unidad de estas fábricas son respectivamente: 0.2, 0.3, 0.2, y 0.5. Calcular el precio medio de venta si se quiere ganar el 20% por unidad. Rp. $1.692/4.7=0.36$

14. El sueldo medio de los obreros de una fábrica es de \$286.

a) ¿Qué porcentajes de hombres y mujeres trabajan en la fábrica si sus sueldos medios respectivos son \$300 y \$260?.

b) Si el 60% de los obreros tienen menos de 30 años y percibe el 20% del total de los sueldos, ¿cuánto es el sueldo medio de los obreros de al menos 30 años?.

Rp. a) 65% y 35%, b) $286N\times0.8/0.4N = 572$\$.

15. En una empresa donde el sueldo medio es de $400 se incrementa un personal igual al 25% del ya existente con un sueldo medio igual al 60% de los antiguos. Si 3 meses más tarde se incrementan cada sueldo en 20%, más 30$, ¿cuánto es el nuevo salario medio?.

Rp. $[510n+318(0.25n)]/1.25n=$ $ 471.6.

16. En el presente mes, 9 vendedores realizaron las siguientes ventas en dólares:

Vendedor	1	2	3	4	5	6	7	8	9
Venta ($)	800	700	500	400	1000	1200	820	750	450

a) ¿Cuánto es la media de las ventas?
b) ¿Quien es el vendedor promedio?.

Rp. a) $735.56, b) Promedio es la mediana=750, que corresponde al vendedor 8.

17. Al tabular las calificaciones de un examen se obtuvieron las siguientes notas: 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 17 y las frecuencias del número de alumnos respectivas: 1, 1, 1, 1, 1, 6, 8, 16, 18, 20, 2.

a) ¿Cuánto es la media, la mediana y la moda de las notas?, ¿qué valor escogería como el promedio?.
b) ¿Cuánto es la nota mínima para estar en el quinto superior?.

Rp. a) $\bar{x}=14.253$, $Me=15$, $Mo=16$, , la Me, b) 16.

18. Los sueldos en una empresa varían de $300 a $800 distribuidos en forma simétrica en 5 intervalos de igual amplitud, con el 15%, 20%, y 30% de casos en el primer, segundo y tercer intervalo respectivamente.

a) Calcule los diferentes indicadores de tendencia central.
b) Si se aplica un impuesto a los sueldos localizados en el cuarto superior, ¿a partir de que sueldo se paga el impuesto?.

Rp. a) $\bar{x}=Me=Mo=550$, b) 650$.

19. A una muestra se aplicó un test para medir autoestima y los puntajes se tabularon en una distribución de frecuencias de 5 intervalos de igual amplitud, siendo la puntuación mínima 25, la tercera marca de clase 62.5. Si las frecuencias en porcentajes del primero al tercero son: 5, 15, 25, y si y el 90% de las puntuaciones son menores que 85

a) calcule el promedio adecuado.
b) y si se considera normal una autoestima comprendida entre 58 y 80 puntos, ¿qué porcentaje de la muestra no tiene autoestima normal?.

Rp. frecuencias: 5, 15, 25, 45, 10, a) $Me=71.67$, b) 50%.

20. En un estudio comparativo del porcentaje de rendimiento de ciertos bonos se elaboró una distribución de frecuencias de 5 intervalos de amplitud iguales, siendo las marcas de clase primera y quinta 15 y 55 respectivamente. Si el 65%

de los bonos rinden menos del 40%, el 25% menos del 30%, el 90% menos del 50% y el 95% al menos 20%,

a) Calcule los promedios de rendimiento.
b) Si el 50% de los bonos de mayor rendimiento deben pagar un impuesto, ¿a partir de que rendimiento corresponde pagar el impuesto?.
c) ¿Es la mediana, el punto medio entre los cuartiles 1 y 3?.

Rp. a) x_{min}=10, A=10, frec. %: 5, 20, 40, 25, 10, $\bar{x}$ =36.5, Me=36.25, Mo=35.7. b) 36.25, c) No,

21. En una prueba de aptitud mental la menor y mayor puntuación fueron 50 y 199 respectivamente. Los puntajes (sin decimales) se tabularon en una distribución de frecuencias simétrica de 5 intervalos de igual amplitud donde el 20% de los casos son menores 95 y el 70% de los casos son menores que 140.

a) Hallar el intervalo centrado en la mediana donde se encuentran el 50% de los puntajes.
b) ¿Es el cuartil 2, el punto medio entre los cuartiles 1 y 3?.

Rp. A = 30, frecuencias 0.1, 0.2, 0.4, 0.2, 0.1, Me = 125, [102.5. 147.5].b) Si.

22. El consumo mensual de agua (en metros cúbicos) de una muestra de 225 viviendas, se tabularon en una distribución de frecuencias simétrica de cinco intervalos de amplitud iguales. Si el consumo mínimo es de 35 m^3, el consumo promedio de 45 m^3, y si 1/3 de la muestra consume al menos 43 m^3 pero menos de 47 m^3.

a) ¿Qué porcentaje de la muestra consume al menos 47 m^3 ?.
b) ¿Cuántos metros cúbicos como mínimo consumen el 60% de las viviendas con mayor consumo?.

Rp. A = 4, a) 75 de 225, b) P_{40}=43.8.

23. Los porcentajes de artículos defectuosos encontrados en un número determinado de cajas recibidas varían de 10 a 25 y han sido tabulados en una distribución de frecuencias simétrica de 5 intervalos de igual amplitud, siendo las frecuencias relativas respectivas del primero al tercero 0.08, 0.24, 0.36. Una caja se considera óptima si el porcentaje de defectuosos no supera el 17% y casi optima si no supera el 20%.

a) Calcular el porcentaje de cajas óptimas y casi óptimas.
b) Si las utilidades por cada caja es de 30 unidades monetarias (u.m.) para las óptimas, 15 u.m. para las casi óptimas y 5 u.m. para el resto, ¿cuánto es la utilidad promedio por caja?.

Rp. a) Porcentajes de cajas: óptimas 44%, casi óptimas: 32%. b) Utilidad promedio=0.44×30+0.32×15+0.24×5=19.2 u.m.

24. Los salarios que ofrece una empresa a los practicantes varían entre $150 y 270$. Si los salarios se agrupan en cuatro intervalos de clase de longitudes iguales de manera que el 40% de los practicantes tienen salarios menores o

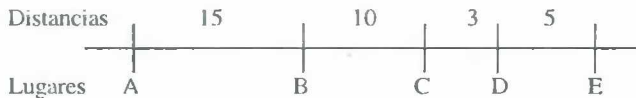
iguales que \$195, el 80% tienen salarios menores o iguales que \$225 y el 15% tiene salarios mayores que \$232.50.

a) ¿Qué porcentaje de practicantes tiene un salario superior al salario medio?.

b) Si el ingreso mínimo se fija en \$240 y la empresa aumenta una misma cantidad a todos los practicantes de modo que el 20% supere el ingreso mínimo, ¿cuánto sería el aumento?, ¿cuánto el salario medio?.

Rp. a) frecuencias: 0.10. 0.60. 0.20. 0.10, media=204\$. 42%. b) \$15. \$219.

25. Cinco personas que viven en los lugares A, B, C, D, y E separadas a las distancias en Km., como se indica en la figura que sigue, deben reunirse en algún lugar.



Determine el lugar de reunión de manera que el costo total del transporte sea mínimo, si el costo de cada transporte es proporcional.

a) al recorrido, b) al cuadrado del recorrido.

Rp. a) a 25 Km. de A b) a 20.2 Km. de A

26. Los pobladores de 6 pueblos: A, B, C, D, E, F ubicados a lo largo de la carretera marginal de la selva y en línea recta, desean construir una escuela. Si la población escolar es: 5% de A, 20% de B a 15 km. de A, 30% de C a 20 km. de B, 20% de D a 15 km. de C, 10% de E a 8 km. de D y 15% de F a 6 km. de E, ¿en qué lugar debe construirse la escuela de manera que

a) el total de las distancias sea mínima?.

b) el costo total del transporte sea mínimo si el costo del transporte es proporcional al cuadrado de la distancia?.

Rp. a) 35km. de A, b) a 38.9 km. de A

27. Un conjunto de *n* artículos cuyos valores de venta serán de \$5, \$7 y 10\$ con las frecuencias respectivas de 20%, 25%, y 55% tienen un costo de producción fijo de \$*k*. Hallar el valor de *k* si se quiere hacer una inversión mínima y si se supone que la inversión es:

a) Es igual a la suma de todas las utilidades.

b) Es igual a la suma de los cuadrados de todas las utilidades.

Rp. a) k=mediana=\$10, b) k=media=$5\times0.2+7\times0.25+10\times0.55$=\$8.25.

Media Geométrica y armónica

28. Si la población de una ciudad fue de 10,000 habitantes en 1950, de 40,000 en 1970 y de 640,000 en 1990, calcule la tasa de cambio promedio.

Rp. $\bar{x}_G$ =8 veces cada 20 años.

29. Un ahorro de 100$ acumula intereses variables de 3%, 5%, 8%, durante 3 años, calcular:

a) El monto del ahorro por año.
b) La tasa promedio del crecimiento del ahorro en los tres años.
c) El porcentaje promedio de crecimiento del ahorro

Rp. a) 103, 108.15, 116.802, b) $\overline{x}_G$=1.05313 , c) 5.31%

30. Si la producción de azúcar en 1991 bajó 20% con relación al año 1990 y si en 1992 aumentó en 20% con respecto a 1991, calcule la tasa promedio del crecimiento de la producción.

Rp. $\overline{x}_G$ =0.9798%, bajó 2%

31. En cuatro meses consecutivos los precios de un artículo fueron $500, $550, $440, y $462 respectivamente, ¿es la tasa de variación promedio igual al −16.7%?., si no es así, ¿cuánto es?. Rp. No, es −2.6%.

32. El crecimiento de la población estudiantil, con respecto al semestre anterior fue como sigue: aumentó 10% en el segundo, aumento 20% en el tercero, y bajó 15% en el cuarto. Encuentre la tasa promedio de crecimiento en los tres semestres.

Rp. media geométrica de: 1.1, 1.2, 0.85, es 3.91%.

33. El rendimiento de 3 automóviles que recorrieron 200 km. cada uno fue de 50, 45, y 60 kilómetros por galón. Hallar el rendimiento promedio de los 3 automóviles. Rp. $\overline{x}_H$ =50.94

34. Tres mecanógrafas escriben 40, 50, y 80 palabras por minuto, si cada una de ellas escribe un mismo texto, calcule la velocidad media.

Rp. $\overline{x}_H$ =52.2

35. Tres obreros utilizaron 480, 360, 240 minutos respectivamente para hacer cierto tipo de objetos. Si utilizaron 0.8, 1, 1.5 minutos por objeto, calcular el tiempo promedio por objeto. Rp. $\overline{x}_H$ =1080/1120=0.96

36. Una estación de servicio automotriz gasta $500 en latas de aceite que cuestan $10 la docena; $500 en latas que cuestan $12.5 la docena; $500 más en latas que cuestan $20 la docena y $500 en otras que cuestan $25 la docena.

a) Determinar el costo promedio por docena de las latas de aceite.
b) En promedio, ¿cuántas docenas se compró?.

Rp. a) 2,000/135 = 14.8148. b) 135/4=33.75

37. Durante los días lunes, martes, miércoles, jueves, y viernes Una persona A compró 70 acciones cada día de la Compañía WWW . Otra persona B invirtió diariamente S/.1800 para comprar acciones de dicha compañía. Si los precios de las acciones cada día fueron como sigue:

Lunes	20
Martes	22.5
Miércoles	24
Jueves	25
Viernes	30

a) Determinar el costo promedio por acción para cada una de las dos personas.
b) ¿Quién consiguió el menor costo promedio por acción?.

Rp. a) A: 8505/350=24.3, B: 9000/377=23.87, b)B

38. Una propiedad de la media dice que si $x_1, x_2, \ldots x_n$ son n valores positivos con media $\bar{x}$ y si la función $f(x)$ es cóncava hacia arriba, entonces $f(\bar{x}) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i)$. Utilizando la propiedad indicada pruebe que $\bar{x}_H \leq \bar{x}$.

donde $\bar{x}_H = \dfrac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$ es la media armónica de los datos.

Capitulo 3

MEDIDAS DE DISPERSIÓN

3.1 Introducción

Las medidas de tendencia central no son suficientes para describir un conjunto de valores de alguna variable estadística. Los promedios determinan el centro, pero nada indican acerca de cómo están situados los datos respecto al centro.

En primer lugar se necesita una medida del grado de *dispersión* o *variabilidad* con respecto al centro con la finalidad de ampliar la descripción de los datos o de comparar dos o más series de datos.

En segundo lugar se necesita una medida del grado de *asimetría* o deformación en ambos lados del centro de una serie de datos, con el fin de describir la forma de la distribución de los datos. Esta medida se denomina *índice de asimetría*.

En tercer lugar se necesita una medida que nos permita comparar el *apuntamiento* o *curtosis* de distribuciones simétricas con respecto a la distribución simétrica normal. Esta medida se denomina *índice de apuntamiento o curtosis.*

Las estadísticas de asimetría y apuntamiento se incluyen en este capítulo dada su poca importancia.

El lector debería correr paquetes de computo entre otros el ***MCEST*** para las aplicaciones de este capitulo.

3.2 Medidas de dispersión

Las *medidas de dispersión* o *variabilidad* son números que **miden el grado de separación de los datos con respecto a un valor central**, que generalmente es la media aritmética.

Las principales medidas de dispersión son:

el rango,

el rango intercuartil,

la varianza,

la desviación estándar, y

el coeficiente de variación.

3.2.1 Rango o recorrido de una variable

Definición. El *rango* de variación o *recorrido, R,* de una serie de datos. es la diferencia entre sus valores máximo y mínimo. Esto es,

$$R = x_{\max} - x_{\min}$$

siendo x_{max} el valor máximo y x_{min} el valor mínimo.

El rango es una medida de dispersión muy fácilmente calculable, pero es muy inestable, ya que depende únicamente de los dos valores extremos. Su valor puede cambiar grandemente si se añade o elimina un sólo dato. Por tanto su uso es muy limitado.

Por **ejemplo,** dadas las dos series de datos

a) 1, 4, 4, 5, 5, 5, 5, 6, 6, 9
b) 1, 2, 3, 4, 5, 6, 7, 8, 9

Ambas series tienen la misma media, 5, y el mismo rango, 8, pero las dos series no tienen la misma dispersión, ya que la segunda tiene mayor variabilidad.

El empleo del rango como medida de comparación de variación puede estar justificado cuando se precise rápidamente de una medida de dispersión y no haya tiempo de calcular algunas de las otras.

3.2.2 Rango intercuartil y rango semiintercuartil

Definición. El *rango intercuartil, RI,* es la diferencia entre sus cuartiles tercero y primero. Esto es,

$$RI = Q_3 - Q_1 .$$

El rango intercuartil **es una medida que excluye el 25% más alto y el 25% más bajo,** dando un rango dentro del cual se encuentra el 50% central de los datos

observados y a diferencia del rango total no se encuentra afectada por los valores extremos.

Si el rango intercuartil es muy pequeño entonces describe alta uniformidad o pequeña variabilidad de los valores centrales.

Por **ejemplo**, si en una distribución de frecuencias de 100 ingresos quincenales se encuentran los cuartiles $Q_1 = 62\$$, y $Q_3 = 70\$$, entonces, el rango intercuartil es $RI = Q_3 - Q_1 = \$70 - 62\$ = 8$. Esto, indica que el 50% de los ingresos quincenales de los 100 empleados varía dentro del valor $8.

El rango intercuartil se aplica a variables medidas en escala por lo menos ordinal.

Definición. El *rango semiintercuartil, RSI,* es igual al rango intercuartil dividido por 2.

El rango semiintercuartil se puede asociar con la mediana y se puede expresar en función de ella. Si una distribución es normal los cuartiles Q_1 y Q_3 son equidistantes de la mediana. Se deduce entonces, que el rango intercuartil y la mediana $\pm$ *RSI*, son la misma distancia. Además, como exactamente el 50% de los datos se encuentran en el rango intercuartil, entonces, el intervalo: *mediana* $\pm$ *RSI* contiene también exactamente el 50% de los datos. Si la distribución es asimétrica, el intervalo: *mediana* $\pm$ *RSI* contendría *aproximadamente* el 50% de los datos.

Por **ejemplo**, si en la distribución de los 100 ingresos quincenales donde $Q_1 = 62\$$, y $Q_3 = 70\$$, el rango semiintercuartil es $4. Si la mediana fuera igual a $66, entonces, aproximadamente el 50% de los datos se hallan comprendidos en el intervalo $66 $\mp$ 4.

NOTA. Si la distribución es muy asimétrica, el rango intercuartil (o el semiintercuartil) es preferible a la desviación estándar como medida de la dispersión.

3.2.3 Varianza y Desviación estándar

La varianza, es una medida que cuantifica el grado de dispersión o de variación de los valores de una variable cuantitativa con respecto a su media aritmética. Si los valores tienden a concentrarse alrededor de su media, la varianza será pequeña. Si los valores tienden a distribuirse lejos de la media, la varianza será grande.

La varianza calculada a partir de una muestra será denotada por s^2 y referida a una población se denotará por σ^2.

Definición. La ***varianza*** se define como **la media aritmética de los cuadrados de las diferencias de los datos** con respecto a su media aritmética.

La varianza es una medida de dispersión con unidades de medición al cuadrado, por ejemplo, $\$^2$, Km^2, etc.

Definición. La ***desviación estándar*** es la raíz cuadrada positiva de la varianza.

La desviación estándar calculada a partir de una muestra se denotará por s y referida a la población por σ. Esto es, $s=\sqrt{s^2}$, $\sigma=\sqrt{\sigma^2}$.

Cálculo de la varianza

1) Varianza de datos no tabulados

La varianza de n valores $x_1, x_2,..., x_n$, de alguna variable cuantitativa X cuya media es $\bar{x}$, es el número:

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

Es fácil verificar que: $\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$

Por lo tanto,

$$s^2=\frac{\sum_{i=1}^{n}x_i^2}{n}-\bar{x}^2$$

EJEMPLO 3.1

Calcular la varianza y la desviación estándar de los 45 ingresos quincenales sin tabular del ejemplo 1.3

SOLUCION

$$n=45 \quad , \quad \sum_{i=1}^{n}x_i=2682\$. \quad \bar{x}=\frac{2682}{45}=59.6. \quad \sum_{i=1}^{n}x_i^2=166{,}244\$^2$$

Luego, la varianza es el número

$$s^2=\frac{\sum_{i=1}^{n}x_i^2}{n}-\bar{x}^2=\frac{166{,}244}{45}-(59.6)^2=142.151\$^2.$$

Mientras, que la desviación estándar es

$$s = \sqrt{s^2} = \sqrt{142.151} = 11.923\$$$

Observar que la varianza está en $\$^2$, mientras que la desviación estándar está en \$.

2) Varianza de datos tabulados

2a) Variable discreta

La varianza de n valores de una variable estadística discreta X que se clasifican en k valores distintos $x_1, x_2, ..., x_k$ con frecuencias absolutas respectivas $f_1, f_2, ..., f_k$, y cuya media aritmética es $\bar{x}$ se calcula utilizando la fórmula:

$$s^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n}$$

Se verifica que $\sum_{i=1}^{k} f_i (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i x_i^2 - n\bar{x}^2$

Por lo tanto,

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \bar{x}^2$$

EJEMPLO 3.2

Calcular la varianza y la desviación estándar del número de hijos de la muestra de 20 familias del ejemplo 1.2.

SOLUCION.

La distribución del ejemplo 1.2 se repite en el cuadro 3.1 donde se ha insertado una columna de productos $f_i(x_i)^2$.

Entonces, $n = 20$, $k = 5$, $\sum_{i=1}^{k} f_i x_i = 44$, $\bar{x} = \frac{44}{20} = 2.2$, $\sum_{i=1}^{k} f_i x_i^2 = 118$

Luego, la varianza es el número

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \bar{x}^2 = \frac{118}{20} - (2.2)^2 = 1.06 \text{ hijos}^2.$$

La desviación estándar es: $s = \sqrt{s^2} = \sqrt{1.06} = 1.03$ hijos.

Cuadro 3.1 Computo de la varianza:
Caso de variable discreta

x_i	f_i	$f_i x_i$	$f_i x_i^2$
0	1	0	0
1	4	4	4
2	7	14	28
3	6	18	54
4	2	8	32
Total	20	44	118

2b) Varianza de datos tabulados por intervalos

La varianza de n valores de alguna variable X, tabulados en k intervalos, con marcas de clases $m_1, m_2, ..., m_k$, frecuencias absolutas respectivas $f_1, f_2, ..., f_k$ y con media $\bar{x}$ es el número:

$$s^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{n}$$

Se puede verificar que:

$$\sum_{i=1}^{k} f_i (m_i - \bar{x})^2 = \sum_{i=1}^{k} f_i m_i^2 - n\bar{x}^2$$

Por lo tanto,

$$s^2 = \frac{\sum_{i=1}^{k} f_i m_i^2}{n} - \bar{x}^2.$$

EJEMPLO 3.3

Calcular la varianza y la desviación estándar de los 45 ingresos quincenales tabulados del ejemplo 1.3.

SOLUCION.

La distribución del ejemplo 1.3 se repite en el cuadro 3.2 donde se ha insertado una columna de productos $f_i m_i^2$.

Cuadro 3.2. Cálculo de la varianza para datos agrupados por intervalos

Ingresos I_i	Marcas m_i	N#.Personas f_i	Productos $f_i m_i$	Productos $f_i m_i^2$
[26,34[	30	1	30	900
[34,42[	38	2	76	2888
[42,50[	46	4	184	8464
[50,58[	54	10	540	29160
[58,66[	62	16	992	61504
[66,74[	70	8	560	39200
[74,82[	78	3	234	18252
[82,90]	86	1	86	7396
Total		45	2702	167764

$$n = 45,\quad k = 8,\quad \sum_{i=1}^{k} f_i m_i = 2702,\quad \bar{x} = \frac{2702}{45} = 60.044,\qquad \sum_{i=1}^{k} f_i m_i^2 = 167.764$$

Luego, la varianza es el número

$$s^2 = \frac{\sum_{i=1}^{k} f_i m_i^2}{n} - \bar{x}^2 = \frac{167.764}{45} - (60.044)^2 = 122.754\ \$^2.$$

La desviación estándar es: $s = \sqrt{s^2} = \sqrt{122.754} = 11.079$ dólares.

Observar que la varianza de los mismos datos no tabulados es $142.151\$^2$.

NOTA (Cálculo de la varianza con frecuencias relativas)

La varianza se calcula también con frecuencias relativas (o porcentajes). En efecto, si se hace $h_i = f_i / n$ en la varianza de datos tabulados, se tiene

$$s^2 = \sum_{i=1}^{k} h_i m_i^2 - \bar{x}^2, \text{ donde } \bar{x} = \sum_{i=1}^{k} h_i m_i$$

NOTA (Varianza poblacional)

La varianza σ^2 de una población finita de N datos $x_1, x_2, ..., x_N$ sin tabular y cuya media es μ, se define por:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

Si formamos todas las muestras posibles de tamaño n y calculamos sus varianzas utilizando la fórmula $s^2 = \sum(x_i - \bar{x})^2 / n$, resulta que la media de todas estas varianzas vale:

$$\frac{n-1}{n}\sigma^2 .$$

Para que la media de todas las varianzas sea igual a σ^2, basta multiplicar a s^2 por $n/(n-1)$. Por esta razón, algunos autores definen la varianza (en estadística descriptiva) con denominador $n-1$. Estas 2 varianzas se tratan en el capítulo 9 de estimación de parámetros.

3.2.4 Coeficiente de variación

Definición. El *coeficiente de variación, C.V.* es una *medida de dispersión relativa* (libre de unidades de medidas), que se define como la desviación estándar dividido por la media aritmética. Esto es,

$$C.V. = \frac{s}{\bar{x}}, \text{ o en } \%$$

El coeficiente de variación se utiliza para comparar la variabilidad de dos o más series de datos que tengan medias iguales o diferentes o que tengan unidades de medida iguales o diferentes (por decir, una serie en kilogramos y otra serie en metros).

Por dar un **ejemplo**, si dos secciones H1 y H2 de matemática I, tienen la misma desviación estándar igual a 14, no podemos concluir que los dos horarios tienen la misma variabilidad. Así mismo, si las desviaciones estándares de H1 y H2 son iguales a 2 y 4 respectivamente no podemos concluir que las notas de H2 son más dispersas que las de H1. La variabilidad depende de las medias de los dos grupos.

Si la media del horario H1 es 16 y la media del horario H2 es 11, los coeficientes de variación respectivos son:

$$C.V_1. = \frac{s_1}{\bar{x}_1} = \frac{14}{16} = 0.875, \text{ o } 87.5\%, \quad C.V_2. = \frac{s_2}{\bar{x}_2} = \frac{14}{11} = 1.27 \text{ o } 127\%$$

Es decir, las calificaciones obtenidas en H1 son más homogéneas o tienen menor variabilidad que las calificaciones del horario H2.

3.2.5 Uso de las medidas de dispersión

La varianza viene expresada en unidades cuadráticas en las que vienen expresados los datos. La desviación estándar viene expresada en las mismas unidades en las que vienen expresados los datos. El coeficiente de variación viene expresada en números abstractos (suprimiendo las unidades en las que vienen expresados los datos).

1) Si dos o más series de datos (observados en el mismo tipo de medición) tienen medias aritméticas iguales (o casi iguales) es más dispersa la serie que tiene mayor medida de variabilidad: Rango, o RI, o s^2, o s, o CV.
Si hay marcada asimetría, es preferible comparar con el rango intercuartil.

2) Si dos o más series de datos, no tienen medias iguales (o casi iguales), o no tienen las mismas unidades de medición, entonces, es más dispersa la serie que tenga mayor coeficiente de variación.

NOTA (Valores estandarizados)

Cuando se necesiten **comparar valores observados** que pertenecen a diferentes distribuciones de datos, las que difieren en su media aritmética o en su varianza, o difieren en el tipo de unidad de medida, entonces se usa el *valor estándar Z* que se define

$$Z = \frac{X - \bar{x}}{s}$$

El lector puede verificar que la variable Z estandariza cualquier media en 0 y cualquier varianza en 1. (Probar que: $\bar{z} = 0$, y $s_Z^2 = 1$)

EJEMPLO 3.4

En una evaluación de Matemáticas e Historia resultan las medias 13 y 17 y las desviaciones estándar 3 y 4, respectivamente. Si un alumno obtiene 14 en Matemáticas y 16 en Historia, ¿en cuál de los dos cursos tiene mejor rendimiento relativo?.

SOLUCION

El hecho de que tenga 16 en Historia y 14 en Matemáticas no significa que tiene mejor rendimiento en Historia.

Se deben calcular los rendimientos relativos con la puntuación estandarizada Z

En Matemáticas $z = \frac{14 - 13}{3} = 0.333$

En Historia $z = \dfrac{16-17}{4} = -0.25$

En consecuencia, tiene mejor rendimiento relativo en Matemáticas.

3.2.6 Propiedades de la varianza.

1) La varianza es un número real no negativo y viene expresada en unidades cuadráticas. Mientras, que la desviación estándar viene expresada en las mismas unidades en las que vienen expresados los datos.

2) Dadas, la media $\overline{x}$ y la varianza s_X^2 de n datos de una variable X, la suma total de los cuadrados de los valores es igual a $n(s_X^2 + \overline{x}^2)$. Para datos no tabulados se tiene por ejemplo,

$$\sum_{i=1}^{n} x_i^2 = n(s_X^2 + \overline{x}^2)$$

3) Si cada uno de n los valores x_i es transformado en $y_i = ax_i + b$, entonces, la varianza de los n valores y_i es, $s_Y^2 = a^2 s_X^2$ (verificar!).

Consiguientemente, $s_Y = |a| s_X$

Como casos particulares se tiene:

Si $y_i = b$, entonces, $s_Y^2 = 0$. Es decir, si los n datos son iguales a una constante, entonces, su varianza es igual a cero.

Si $y_i = x_i + b$, entonces, $s_Y^2 = s_X^2$. Es decir, si sumamos a cada dato una constante, la varianza (y la desviación estándar) no cambian.

Si $y_i = ax_i$, entonces, $s_Y^2 = a^2 s_X^2$. Es decir, si multiplicamos a cada dato por una constante, a, la varianza de los nuevos valores es igual que la varianza de los antiguos valores multiplicada por a^2.

4) La varianza y la desviación estándar pueden ser calculadas también en distribución de frecuencias de intervalos de amplitud diferentes, siempre que puedan determinarse las marcas de las clases. Por otra parte, dependen de todos los datos y son sensibles a la variación de cada uno de estos. Basta que uno de los datos varíe, para que varíen aquellas.

5) Dados k series de datos con tamaños, medias y varianzas respectivas $n_1, \bar{x}_1, s_1^2$, $n_2, \bar{x}_2, s_2^2$, ..., $n_k, \bar{x}_k, s_k^2$, entonces, la varianza, s_T^2, de los $n_1 + n_2 + ... + n_k = n$ datos es:

$$s_T^2 = \frac{\sum_{i=1}^{k} n_i s_i^2}{n} + \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{n}.$$

6) **Desigualdad de Chebyshev**.

Cualquiera sea la forma de la distribución de frecuencias (simétrica o asimétrica), el intervalo $[\bar{x} - ks_X, \bar{x} + ks_X]$, $k > 1$ contiene por lo menos el

$$1 - \frac{1}{k^2} \text{ en \% de los datos.}$$

El porcentaje de datos que se hallan fuera del intervalo es menor que el

$$\frac{1}{k^2} \text{ en \%}.$$

Por **ejemplo**, el intervalo $[\bar{x} - 2s_X, \bar{x} + 2s_X]$ contiene por lo menos el $1 - \frac{1}{2^2} = \frac{3}{4}$ o 75%, de los datos.

El intervalo $[\bar{x} - 3s_X, \bar{x} + 3s_X]$ contiene por lo menos el 88.89%, (8/9), de los datos.

El intervalo $[\bar{x} - 4s_X, \bar{x} + 4s_X]$ contiene por lo menos el 93.75%, (15/16), de los datos.

EJEMPLO 3.5

En el mes de Enero el sueldo promedio de los trabajadores del sector industrial era de \$200. Para el mes de Julio se considera un aumento del 30% al sueldo del mes de Enero más un adicional de \$50. Si el coeficiente de variación en Enero era de 0.25, ¿se puede decir que la distribución de sueldos en Julio es más homogénea?

SOLUCION

Sea X: Sueldos de Enero, Y: Sueldos de Julio

La media de Enero es: $\bar{x} = \$200$.

Coeficiente de variación en Enero, $CV = 0.25$

La desviación estándar de Enero es $s_X = CV \times \bar{x} = 0.25 \times 200 = \$50$

La relación entre las dos variables es:

$$Y = 1.30X + 50,$$

Entonces, la media de los sueldos de Julio es

$$\bar{y} = 1.3\bar{x} + 50 = 1.3(200) + 50 = 310$$

La varianza de los sueldos de Julio es

$$s_Y^2 = (1.3)^2 s_X^2 = (1.3)^2(50)^2 = 4225$$

La desviación estándar: $s_Y = \sqrt{4225} = 65$

Coeficiente de variación en Julio: $CV = \dfrac{s_Y}{\bar{y}} = \dfrac{65}{310} = 0.2097.$

Comparando los coeficientes de variación de Enero y Julio se puede decir que la distribución de los sueldos de Julio es más homogénea.

EJEMPLO 3.6

Si el ingreso de 120 obreros tiene una media de \$300 y una desviación estándar de \$30

a) ¿Cuántos obreros por lo menos tienen sueldos comprendidos en el intervalo [\$240, \$360]?.
b) Determinar el intervalo que contiene al menos el 88.889% de los ingresos
c) Si el mínimo sueldo es \$210, en qué porcentaje se puede afirmar que los ingresos son superiores a \$390?

SOLUCION.

a) $\bar{x} = \$300$, $s = \$30$, de la relación

$$[300 - k(30),\ 300 + k(30)] = [240,\ 360]$$

resulta $k = 2$. Entonces el, $1 - \dfrac{1}{2^2} = \dfrac{3}{4}$ o 75%(120)=90 obreros por los menos tienen ingresos en el intervalo [240, 360].

b) Si al menos el 88.889% de los obreros tienen ingresos en el intervalo

$$[300 - k(30),\ 300 + k(30)]$$

entonces, $1 - \dfrac{1}{k^2} = 0.8889$. De donde resulta $k = 3$.

Luego, el intervalo es [\$210, \$390].

c) Fuera del intervalo [\$210, \$390] está menos del 11.11% de los ingresos. Si el mínimo es \$210, entonces, el porcentaje de ingresos mayor que \$390 es **menos** de 11.11%.

EJEMPLO 3.7

El costo inicial de producción, X ; de una muestra de 80 objetos de cierto tipo, tiene una desviación estándar de \$30. La media del costo de producción es de \$250 para el 60% de la muestra y de \$200 para el resto. El costo final de producción Y es dado por la relación:

$$Y = 1.2X + 5.$$

Si el precio de venta de cada objeto de la muestra es proporcional al cuadrado del costo final de producción, ¿cuánto se recaudaría por la venta total?.

SOLUCION.

$s_X = \$30, \quad \bar{x} = 250\times0.60 + 200\times0.40 = \$230$

De $Y = 1.2X + 5$, se tiene, $\bar{y} = 1.2\bar{x} + 5 = 1.2(230) + 5 = 281$. También,

$s_Y^2 = (1.2)^2 s_X^2 = (1.2)^2(30)^2 = 1296$

Recaudación total $= \sum_{i=1}^{80} y_i^2 = 80(s_Y^2 + \bar{y}^2) = 80(1296+(281)^2) = 6{,}420{,}560.$

3.3 Indices de asimetría

Definición. Se dice que una distribución de frecuencias es *simétrica*, si los intervalos equidistantes del intervalo central tienen iguales frecuencias. También se dice que una distribución es simétrica si su curva de frecuencias es simétrica con respecto al centro de los datos.

Dos distribuciones pueden tener la misma media y la misma desviación estándar, pero pueden diferir en el grado de asimetría.

Si la distribución es simétrica, entonces, la media, la mediana y la moda coinciden. En contraposición, si estos 3 promedios no coinciden la distribución tiene que ser asimétrica.

Existen varias medidas de asimetría, una de ellas es el *coeficiente o índice de asimetría de Pearson.*

Definición. El *índice de asimetría de Pearson* es el número

$$As = \frac{\bar{x} - Mo}{s}$$

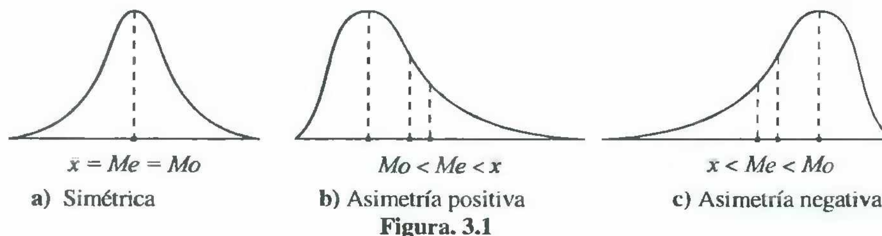
Como en distribuciones asimétricas se verifica : $\bar{x} - Mo \cong 3(\bar{x} - Me)$, entonces, otra forma de expresar el índice de asimetría es:

$$A_S = \frac{3(\bar{x} - Me)}{s}.$$

Interpretación.

Si la distribución de los datos es simétrica, $As = 0$. Ver la figura 3.1.a, donde se observa, además, que coinciden los tres promedios: $\bar{x} = Me = Mo$.

Si $As \neq 0$, la distribución es asimétrica. Además, es **asimétrica positiva** o sesgada a la derecha, si $As > 0$, (Fig. 3.1 b donde $Mo < Me < \bar{x}$). Y, es asimétrica negativa o sesgada a la izquierda si $As < 0$ (Fig. 3.1.c donde $\bar{x} < Me < Mo$)



a) Simétrica **b)** Asimetría positiva **c)** Asimetría negativa

Figura. 3.1

Por **ejemplo**, la distribución de los 45 ingresos quincenales del ejemplo 1.3 tabulados en ocho intervalos tiene asimetría negativa:

$$As = \frac{3(\bar{x} - Me)}{s} = \frac{3(60.44 - 60.75)}{11.079} = -0.191$$

NOTA. (Otros índices de asimetría)

El índice **de asimetría de Pearson** utilizando momentos es definido por:

$$As = \frac{nM_3}{(n-1)(n-2)s^3}$$

donde $M_3 = \sum_{i=1}^{n}(X_i - \bar{x})^3$, $n =$ número de casos, $s =$ la desviación estándar.

Este índice es utilizado por los paquetes de computo estadístico para determinar la asimetría de distribuciones de la forma dato-frecuencia.

Para ***n*** **datos tabulados en** ***k*** **intervalos**, un método alternativo es utilizar el índice de **asimetría de Fisher** definido por:

$$As = \frac{M_3 / n}{s^3}$$

donde: $M_3 = \sum_{i=1}^{k} f_i (m_i - \bar{x})^3$, $s =$ la desviación estándar

Si la distribución es simétrica $As = 0$. Si $As > 0$, es asimétrica positiva y si $As < 0$, es asimétrica negativa.

Por **ejemplo**, continuando con el ejemplo 1.3, el índice de asimetría de los 45 ingresos quincenales tabulados en la forma dato-frecuencia es $As = -0.375$. Y de los mismos datos tabulados en 8 intervalos es: $As = -0.3$.

NOTA (Ojivas asimétricas y simétricas). Las ojivas o curvas de frecuencias acumuladas. presentan formas particulares según el tipo de asimetría. Por ejemplo, en la figura 3.2a la curva de frecuencia acumulada A es de una distribución con asimetría extrema negativa. La Ojiva C es de asimetría extrema positiva. La ojiva B es de una distribución simétrica. En la figura .3.2b la diagonal B es la ojiva de una distribución normal. La curva F es la ojiva de una distribución simétrica leptocúrtica, y la E de una platicúrtica. (ver 3.4 curtosis)

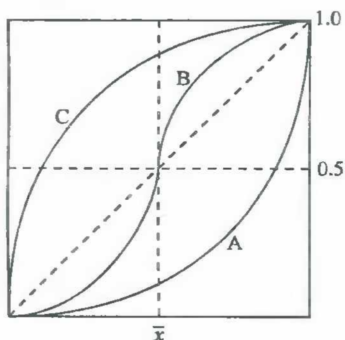


Fig. 3.2a Ojivas asimétricas relativas

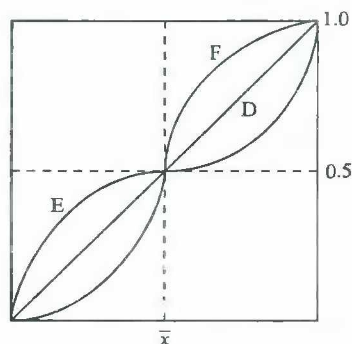


Fig. 3.2b Ojivas simétricas relativas

3.4 Curtosis

La curtosis es la propiedad de una distribución de frecuencias por la cual se compara la dispersión de los datos observados cercanos al valor central con la dispersión de los datos cercanos a ambos extremos de la distribución. La curtosis se mide en comparación a la curva simétrica normal o *mesocúrtica* (fig. 3.3a)

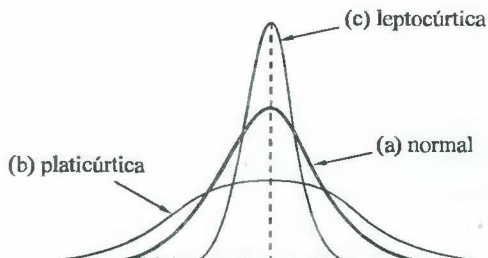


Fig. 3.3 Curtosis de curvas simétricas

Una curva simétrica con curtosis mayor que de la normal es denominada curva *leptocúrtica* (fig. 3.3c).

Una curva simétrica con curtosis menor que de la normal es denominada curva *platicúrtica* (fig. 3.3b).

Existen varias maneras de medir la curtosis de la distribución de los datos.

Curtosis basado en percentiles

Esta medida de curtosis es muy poco usada por ser muy inestable. Sin embargo. describe muy bien el concepto.

En una curva normal. el cociente del rango intercuartil (percentil 75 menos el percentil 25) entre la diferencia del percentil 90 menos el percentil 10 es aproximadamente igual 0.5. A medida que $P_{75} - P_{25}$ y $P_{90} - P_{10}$ sean iguales (valor del cociente casi uno), la distribución será leptocúrtica, y a medida que $P_{75} - P_{25}$ sea cada vez más pequeño con respecto a $P_{90} - P_{10}$ (valor del cociente casi cero) la distribución será platicúrtica.

La curtosis utilizando percentiles se define por el cociente:

$$K = \frac{P_{75} - P_{25}}{P_{90} - P_{10}} - 0.5$$

Interpretación. Si la distribución es **normal**, K tiende a 0. Si K tiende a 0.5, es **leptocúrtica**, y si K tiende a –0.5, es **platicúrtica**.

Por **ejemplo**, la distribución de los 45 ingresos quincenales del ejemplo 1.3 tabulados en 8 intervalos tiene curtosis $K = (66.75 - 53.4)/(73.5 - 45) - 0.5 = -0.03$. Sin embargo, no se puede relacionarla con una distribución normal, por que ésta distribución de frecuencias no es simétrica.

NOTA. (Otras medidas de curtosis)

La curtosis utilizando momentos es definida por la expresión:

$$K = \frac{n(n+1)M_4 - 3M_2M_2(n-1)}{(n-1)(n-2)(n-3)s^4}$$

donde $M_j = \sum_{i=1}^{n}(X_i - \bar{x})^j$, $n =$ número de casos, $s =$ la desviación estándar.

Esta curtosis es utilizado por los paquetes de computo estadístico para determinar la curtosis de **distribuciones de la forma dato-frecuencia.**

Para ***n* datos tabulados en *k* intervalos**, la curtosis se calcula por:

$$K = \frac{M_4 / n}{s^4} - 3$$

donde: $M_4 = \sum_{i=1}^{k} f_i (m_i - \bar{x})^4$, s = la desviación estándar.

Si la distribución es **normal**, K=0. Si K>0, es **leptocúrtica**, y si K<0 es **platicúrtica**.

Por **ejemplo**, continuando con el ejemplo 1.3, la curtosis de los 45 ingresos quincenales tabulados en la forma dato-frecuencia es $K = 1.021$. Y de los mismos datos tabulados en 8 intervalos es $K = 0.244$. **Sin embargo, no se puede decir que es leptocúrtica, por que la distribución de los datos no es simétrica.**

3. 5 Diagramas de caja

Existe Una gran variedad de gráficas estadísticas para extraer información acerca de las propiedades de un conjunto de datos.

Una gráfica útil para reflejar propiedades de los datos es la **gráfica de caja** ("**box plots**") que se basa en la mediana (o en la media), los cuartiles y valores extremos. La caja representa el rango intercuartil que encierra el 50% de los valores y tiene la **mediana** (Me) dibujada dentro. El rango intercuartil tiene como extremos el percentil 75, P_{75} (cuartil superior) y el percentil 25, P_{25} (cuartil inferior).

Además de la caja se incluye la **extensión** de los datos mediante segmentos que se extienden de la caja hacia el valor máximo (U) y hacia el valor mínimo (L) de los datos. Este recuadro se dibuja con el eje de la variable en forma horizontal o vertical como se indica en la figura que sigue.

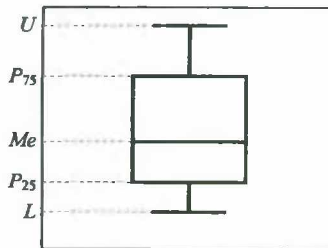


Diagrama de Caja y Extensiones

De un gráfico de cajas, se obtiene información de los datos acerca de:
La centralización (Observando la ubicación de la mediana)
La dispersión o variabilidad (mediante el rango intercuartil: $RI = P_{75} - P_{25}$)
La asimetría (comparando: $Me - P_{25}$ con $P_{75} - Me$)
Las colas (por la longitud de los segmentos que salen de los lados de la caja)

EJEMPLO 3.8.

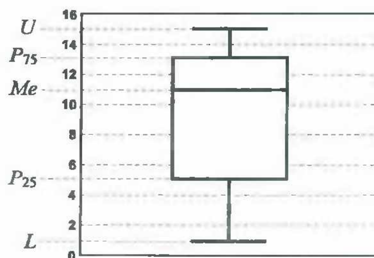
En una prueba de conocimientos 20 alumnos han obtenido las notas:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 12, 13, 14, 15, 12, 13, 14, 15

analizar los datos utilizando una gráfica de caja.

SOLUCION

Se observa que la mediana es igual a 11. El percentil 75 es 13 y el percentil 25 es 5.5. La extensión superior es el dato máximo 15 y la extensión inferior es el dato mínimo 1. Además, la distribución de los datos tiene asimetría negativa.



Los **datos atípicos o discordantes o raros llamados "outliers"** (***aislados***) son aquellos que se ubican fuera del intervalo $[P_{25}-1.5RI,\ P_{75}+1.5RI]$, donde $RI = P_{75} - P_{25}$ es el rango intercuartil. En este caso, el extremo inferior L es el dato **mínimo no outlier** y el extremo superior U es el dato **máximo no outlier**.

Existen muchos paquetes de computo que proporcionan las gráficas de caja.

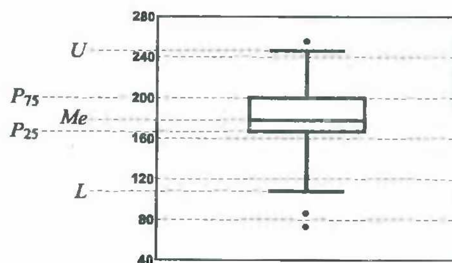
EJERCICIO.

Los ingresos quincenales en dólares de 40 familias se registraron de la siguiente manera:

109, 174, 158, 211, 164, 179, 137, 175,
192, 147, 203, 186, 072, 246, 193, 163,
231, 197, 170, 190, 169, 188, 140, 237,
179, 085, 217, 168, 185, 208, 164, 175,
228, 124, 255, 151, 182, 167, 209, 169.

analizar los datos utilizando un diagrama de caja.

SOLUCION.



La mediana $Me = 177$.
El percentil 75 es 200 y el percentil 25 es 163.5. Rango Intercuatril: d =36.5.
$P_{75} + 1.5d = 200 + 1.5\times36.5 = 254.75$. entonces. $U = 246$ (máximo no outlier).
$P_{25} - 1.5d = 163.5 - 1.5\times36.5 = 108.9$, entonces, $L = 109$ (mínimo no outlier).
Los outlieres (datos aislados) son: 255 en el lado superior y 72 , 85 en el lado inferior. La distribución tiene asimetría positiva.

Los diagramas de caja se usan también para **comparar la variabilidad** de dos o más grupos de datos cuando están expresados en la misma unidad de medida.

Además se usan para **comparar medias poblacionales.** (Si dos cajas no se traslapan, entonces. la diferencia entre las medias es significativa. Ver referencia [15] página 336).

EJERCICIOS

1. A cuatro unidades estadísticas se le asigna los valores 6, 10, 14, y 20 respectivamente en una escala de razón. Si en la misma escala se transforma 6 en 9, calcular el coeficiente de variación de los 4 valores transformados.

Rp. 7.758/18.75 = 0.41376.

2. En una industria el jornal diario de sus obreros tiene una media de $10 y una desviación estándar de $2. Si se hace un incremento de 20% en cada jornal y una bonificación adicional de $3, ¿en qué porcentaje cambió la variabilidad de los jornales?.

Rp. CV1= 0.2, CV2 = 0.16, bajó en 20%.

3. La distribución de las notas resultantes en un examen de conocimientos tiene media igual a 10, mediana igual a 8, moda igual a 4 y desviación estándar igual a 3. Además, se sabe que el 25% de los alumnos tienen como máximo 02 y el 75% de los alumnos tienen como máximo 11

a) Describa y calcule la asimetría de la distribución
b) Determine la dispersión de la distribución mediante el rango intercuartil
c) Si a cada alumno se sube 4 puntos, ¿se ha logrado bajar la dispersión de las notas? Rp. a) Cola derecha A_S=3, b) RI=11−02=9, c) Si CV_1=0.3, CV_2=0.21

4. Se realizan 10 mediciones con cada uno de dos termómetros A y B. Las medias aritméticas de las medidas es 38 grados centígrados en cada caso y los coeficientes de variación son 1% y 2% respectivamente, ¿cuál de los termómetros es más confiable?.

Rp. CVA < CVB, entonces A es más confiable

5. La media y la desviación estándar de los sueldos de N empleados de una fábrica son 500 y 30 respectivamente. A cada uno de los N empleados se les dará un aumento de $A\%$ de su sueldo más una bonificación de B soles. Halle A y B de tal manera que la media de los sueldos modificados sea 600 y su desviación estándar 33. Rp. $A=0.1$, $B=50$.

6. Un investigador califica a dos grupos A y B que dan una prueba de aptitud, asignándoles valores en escala ordinal. Si el cuartil 1 de A es 5 y de B es 35 y si el cuartil 3 de A es 30 y el de B es 50, ¿cuál de los dos grupos tiene aptitud más homogénea?

Rp. usando el rango intercuartil, B es más homogéneo.

7. Una prueba de conocimientos, A, se calificó sobre 20 puntos dando una media de 12 y una desviación estándar de 2 puntos. Mientras que una prueba de aptitud, B, se calificó sobre 100 puntos, dando una media de 70 y una desviación estándar de 5.

a) ¿En cuál de las dos pruebas los puntajes son más homogéneos?

b) Si Juan tiene 14 en A y Luis 73 en B, ¿quién tiene mejor rendimiento?

Rp. a) B, $C.V_A = 0.167$, $C.V_B = 0.071$, b) Juan tiene mejor puntuación estándar.

8. Los sueldos de 100 empleados de una empresa tienen una media de 300$ y una desviación estándar de $50. Se proponen dos alternativas de aumento: i) $75 a cada uno, ii) 15% del sueldo más $20 a cada uno. ¿Cuál alternativa es más conveniente,

a) Si la empresa dispone sólo de $37,000 para pagar sueldos?.

b) Si la empresa quiere homogeneizar los sueldos?.

Rp. a) total a pagar: i) $100\times375=37,500$, ii) $100\times365=36,500$, conviene ii)
b) $CV1=50/375=0.133$, $CV2=57.5/365=0.156$, conviene i.

9. Los sueldos de 150 trabajadores de una empresa tienen un coeficiente de variación del 5% en el mes de agosto. Para el mes de setiembre hay un aumento a cada trabajador del 20% de su sueldo más una bonificación de $60 y el coeficiente de variación baja a 4%.

a) Calcule la media y la desviación estándar de los sueldos del mes de agosto.

b) ¿Cuánto dinero adicional necesita la empresa para pagar todos los sueldos del mes de setiembre?.

Rp. a) $\bar{x} = 200$, $s = 10$, b) 15,000.

10. La media del salario mensual que pagaba una empresa a sus empleados fue en junio de $300. En julio se incorporó un grupo de empleados igual al 20% de los que habían en junio y con un salario medio igual a $210. En agosto la empresa concedió un aumento general del 15% de los salarios más $30.

a) Calcule el salario medio de todos los empleados en el mes de agosto.

b) Si en julio el coeficiente de variación fue 0.04, ¿cómo ha variado este coeficiente en el mes de agosto con respecto a julio?.

Rp. a) 357.75, b) 13.11/357.75=0.037, bajó.

11. Al calcular la media y la desviación estándar de 80 datos, resultaron 30 y 4 respectivamente. Un chequeo mostró que en lugar del valor 1.7 se introdujo 17. Corrija la media y la desviación estándar.

Rp. $\bar{x} = 29.809$, $s = 4.885$

12. El costo C en dólares por operación en una clínica depende del tiempo X, en horas, que ésta dure, y es igual a:

$$C = 50 + 100x + 250x^2$$

Calcule el costo medio de 30 operaciones si tuvieron una media y una desviación estándar igual a 2 horas.

Rp. $50+100\times 2+250(2^2+2^2)=2250$.

13. La varianza de n, $(n > 4)$, datos de variable X es 40. Si la suma de los datos es 40 y la suma de sus cuadrados es 560, calcular el coeficiente de variación de los datos después de la transformación: $Y = (3X + 9)/10$.

Rp. $n=10$, $s_X=40$, $\bar{x}=4$, $\bar{y}=2.1$, $s_Y=1.897$, $CV_Y=1.897/2.1=0.9$

14. El costo de producción X de una muestra de cierto tipo de objeto tiene una desviación estándar de \$30. El costo medio de producción es de \$250 para el 60% de la muestra y de \$200 para el resto. Si su precio de venta en dólares es dado por la relación $Y = 1.1X + 10$, calcule la media y la varianza de la venta de la muestra. Rp. 263\$ y 1089\$^{2}.

15. El costo inicial de producción, X; de una muestra de 50 objetos de cierto tipo, tiene una desviación estándar de \$3. La media del costo de producción es de \$25 para 30 de los objetos de la muestra y de \$20 para el resto. El costo final de producción Y es dado por la relación:

$$Y = 1.15X + 2.$$

Suponga que el precio de venta de cada objeto de la muestra es proporcional al cuadrado del costo final de producción, ¿cuánto se recaudaría por la venta total de los 50 objetos?.

Rp. $\bar{x}=23$, $\bar{y}=28.45$, $s_Y^2=11.9025$, recaudación=41065.25.

16. En una prueba de aptitud aplicada a 100 personas se obtuvo la siguiente información: Los puntajes se tabularon en una distribución de frecuencias de 5 intervalos de amplitud iguales, siendo el puntaje mínimo, 40 y el máximo, 90.

La frecuencia absoluta del intervalo central fue de 40 y del quinto de 10. La frecuencia relativa del primer intervalo fue de 0.05 y la del cuarto de 0.15.

a) Calcule los cuartiles 1, 2, 3 y utilizando estas medidas analice la asimetría.
b) Calcule la varianza si a cada persona se bonifica con 10 puntos.

Rp. a) Q_1 =56.67, Q_2 =63.75, Q_3 =80, asim positiva, b) s^2=104.75, no cambia si se suma 10.

17. Los siguientes datos muestran los calificativos de 20 personas sometidos a una prueba de aptitud. Los 20 estudiantes fueron divididos en dos grupos, al grupo1 se calificó de 0 a 100 y al grupo 2 de 0 a 20:

Grupo 1: 86, 81, 79, 73, 95, 86, 94, 90, 86, 88.
Grupo 2: 16, 19, 13, 20, 14, 16, 19, 18, 17, 15.

a) Calcule la media y la desviación estándar en cada grupo, ¿cuál de los grupos es más homogéneo?.
b) ¿Se puede aceptar que el estudiante con 73 puntos del grupo 1 tiene mayor aptitud que el estudiante con 13 puntos del grupo 2?.

Rp. a) CV_1=0.0744, CV_2=0.1313, b) No, z_1=–2, z_2 =– 1.67 el del segundo grupo está mejor.

18. Los sueldos en dólares de los empleados de dos empresas A y B se dan en la siguiente tabla de frecuencias:

Sueldos	[50, 90[	[90, 130[	[130, 170[	[170, 210[	[210, 250]
Empresa A	12	14	16	60	20
Empresa B	30	80	15	14	13

a) Calcule la asimetría de las distribuciones A y B. Grafique las ojivas relativas. ¿Es el rango intercuartil de A, menor al rango intercuatil de B?.
b) ¿En qué empresa los sueldos son más homogéneos?
c) Si un empleado de A y otro de B ganan cada uno $130, ¿quien de ellos está mejor ubicado en su centro de trabajo?

Rp. a) AS_A=–0.813, RI_A=61.75, AS_B=–1.019, RI_B=46.667, NO, b) En A,
CVA=47.286/170.328=0.278, CVB =45.907/123.648=0.371,
c) Z_A= –0.853, Z_B=0.1376 el de A

19. Las notas de un examen se tabularon en una distribución de frecuencias de cuatro intervalos de amplitud iguales a cuatro, siendo el dato mínimo igual a cuatro y las frecuencias relativas primera y tercera respectivamente 0.15 y 0.35. Calcule la varianza de la distribución si la media aritmética es 12.4.

Rp. 15.04.

20. Los sueldos en dólares de 50 empleados de una empresa se dan en la siguiente tabla:

Sueldos	[60, 100[	[100, 140[	[140, 180[	[180, 220[	[220, 260]
Empleados	8	10	20	7	5

Se plantean dos alternativas de aumento: La primera, consiste en un aumento general de $50. La segunda consiste en un aumento general del 30% del sueldo, además una bonificación de $10.

a) ¿Cuál de las dos propuestas conviene a los trabajadores si el interés es
   a1) subir la media de los sueldos?, a2) bajar la dispersión de los sueldos?.

b) ¿Es la mitad inferior de los sueldos más homogénea que la mitad superior?

Rp. a1) medias, 1era: $202.8, 2da: 208.64, conviene la 2da, a2) s_1=46.434, s_2=60.3642, CV1=0.229, CV2=0.289, conviene la 1ra. b) No, CV1=0.229, CV2=0.148

21. Un conjunto habitacional está formado por 3 edificios de departamentos. Se tiene los siguientes datos respecto al consumo mensual de electricidad de cada uno de los edificios.

Edificio 1: Tiene 8 departamentos, la media y la desviación estándar de los consumos es S/. 85 y S/.12 respectivamente.

Edificio 2: Tiene 9 departamentos cuyos consumos en soles son: 88, 92, 106, 110, 93, 102, 91, 94, 80.

Edificio 3: Los consumos se dan en la siguiente tabla:

Consumo en soles	Departamentos
[50, 60[	1
[60, 70[	2
[70, 80[	4
[80, 90]	3

a) ¿Cuál de los edificios tiene el menor consumo de electricidad?

b) ¿Cuál es el consumo promedio en todo el conjunto habitacional?

c) ¿En cuál de los edificios los valores que representan los consumos están más dispersos?. Rp. a) Edif 1: con S/.680, b) S/. 84.296, c) en Edifi 1, CV1=0.14.

22. En una empresa el coeficiente de variación de los ingresos de 150 empleados es 69%. Después de un aumento de los sueldos en S/ 85 a cada uno de ellos, el coeficiente de variación resultó 54%. La empresa fija un salario mínimo de S/ 350 lo que beneficia a 50 empleados que antes del reajuste ganaban menos de S/280 con un sueldo promedio de S/ 250. ¿Qué cantidad de dinero necesita mensualmente la empresa para pagar los sueldos después del reajuste?.

Rp. Antes del reajuste la media de los 150 es S/.306, de los 50 es S/.250, y de los 100 es S/. 334. Después del reajuste la media de los 50 es realmente $335 pero sube a S/.350 y de los 100 es S/. 419. Se necesita: $50 \times 350 + 100 \times 419 = 59{,}400$ S/..

23. En una empresa donde trabajan hombres y mujeres la media general de los sueldos es $250. Si la media y la desviación estándar de los sueldos en el grupo de varones es $270 y $15 y en el grupo de mujeres es $220 y $10,

a) Calcule el porcentaje de hombres y mujeres,

b) Calcule la desviación estándar de los sueldos de todos los trabajadores de la empresa. Rp. a) 60% hombres, 40% mujeres, b) $27.84.

24. Las calificaciones de 120 personas que rinden una prueba de aptitud divididos en dos grupos A y B tienen una media total de 208 y una varianza de 1728.6. La media y la varianza de las calificaciones del grupo A es 240 y 225 respectivamente. Si 72 de tales personas forman el grupo A, calcule la media y la varianza de las calificaciones del grupo B. Rp. 160, 144

25. Un producto que proviene de dos fábricas A y B se clasifica en tres clases según su duración: De 1era, si su vida útil está en el cuarto superior, de 3era, si su duración está en el cuarto inferior, en otro caso son de 2da clase. Los precios son los mismos en cada marca A y B y en cada clase. Si A y B tienen medias iguales a 12 meses, 1er cuartil 10 y 8 meses, y si sus curvas de frecuencias son simétricas leptocúrtica y platicúrtica respectivamente, ¿cuál sería su estrategia de compra para adquirir las 3 clases del producto?.

Rp. 1era de B, 3era de A y 2da de A por tener menor rango intercuartil.

26. Los precios de un producto en las 50 tiendas del centro de una ciudad A varían entre 8 y 18 soles. Estos precios se han organizado en una distribución de frecuencias con 5 intervalos de amplitud iguales, resultando que en el 16, 56, 76 y 90 por ciento de estas tiendas los precios fueron inferiores a 10, 12, 14, y 16 soles, respectivamente. Un estudio similar mostró que en las tiendas del centro de otra ciudad B, la media de los precios del mismo producto resultó ser 13.5 soles con una desviación estándar de 3 soles. Una tienda, que tiene sucursales en los centros de las ciudades A y B, vende el producto en la ciudad B a 12 soles. Si esta tienda, tiende a fijar sus precios de acuerdo al medio, estime el precio al que vende este producto en la ciudad A.

Rp. A=2, frec: 8, 20, 10, 7, 5, media =12.24, s=2.3963, Z en B es –0.5 y Z en A es $(x-12.24)/2.3963$, haciendo iguales resulta x=11.04

27. (*Desigualdad de Chebyshev*).

Si $\bar{x}$ y s son la media y la desviación estándar de n datos, pruebe que el intervalo: $[\bar{x}-ks, \bar{x}+ks]$, $k>1$ contiene al menos el $(1-1/k^2)\times 100\%$. de los n datos. Esto es, el número de datos de *cualquier tipo de distribución* que caen en tal intervalo no puede ser menor a $(1-1/k^2)\times 100\%$.

SOLUCION. Sean n_1 y n_2 # de datos dentro y fuera respectivamente del intervalo, entonces, $(n_1+n_2)/n=1$. Por otra parte, la suma de cuadrados total es igual a la suma de cuadrados dentro más suma de cuadrados fuera. Luego,

$$s^2 = \frac{\sum_n (x_i-\bar{x})^2}{n} = \frac{\sum_{n_1}(x_i-\bar{x})^2 + \sum_{n_2}(x_i-\bar{x})^2}{n} \geq \frac{\sum_{n_2}(x_i-\bar{x})^2}{n}$$

Si los datos x_i están fuera del intervalo, entonces, $|x_i-\bar{x}| > ks$, y,

$$SCF = \sum_{n_2}(x_i-\bar{x})^2 > \sum_{n_2} k^2 s^2 = k^2 s^2 n_2$$

Sustituyendo, este último en s^2, resulta, $k^2 s^2 n_2 > ns^2$. Luego, la proporción de datos fuera del intervalo n_2/n=1 es menor que $1/k^2$, y la proporción de datos dentro es al menos $1-1/k^2$.

28. Los tiempos que emplearon 900 personas para hacer una tarea tienen una media de 30 minutos y una desviación estándar de 3 minutos.

a) ¿Cuántas de estas personas por lo menos, emplearon entre 24 y 36 minutos para hacer esa tarea?

b) Determine un intervalo más corto en el que se encuentren los tiempos empleados por al menos 800 de estas personas.

c) Si la distribución de los tiempos empleados es simétrica, ¿qué porcentaje de estas personas emplearon más de 39 minutos?.

Rp. a) k=2, al menos 675 b) $1-k^{-2}=8/9$, k=3, [21, 39], c) menos de 1/18 o 5.55%.

29. Los puntajes obtenidos en una prueba de conocimientos tienen una meda igual a 8. Si el coeficiente de variación de los puntajes es igual 0.25.

a) Halle el porcentaje de evaluados cuyos puntajes estén comprendidos en el intervalo [04, 12].

b) Si el puntaje mínimo es igual a 02, ¿que proporción de los evaluados tienen más de 14?.

Rp. s=3, a) K=2, al menos 75%, b) K=3, al menos 88.89% en [2,14], menos de 11.11%.

30. Pruebe que si $x_1, x_2, \ldots x_n$ tienen media $\bar{x}$ y desviación estándar s, entonces, $\frac{x_1-\bar{x}}{s}, \ldots, \frac{x_n-\bar{x}}{s}$ tiene media cero y desviación estándar 1.

31. Verificar que si dos series de datos tienen respectivamente tamaños: n_1, n_2, medias $\bar{x}_1, \bar{x}_2$ y varianzas: s_1^2, s_2^2, entonces, la varianza (total) de las dos series en conjunto está dada por:

$$s_t^2 = \frac{(n_1 s_1^2 + n_2 s_2^2) + \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^2}{n_1+n_2}.$$

32. La tabla que se presenta a continuación corresponde a un número de personas que se encontró en una muestra tomada en 4 distritos y que son consumidores de un producto. La tabla muestra la clasificación por distrito y por edad y sexo:

Distrito	Edad Hombres			Edad Mujeres		
	20 – 30	30 – 40	40 – 50	20 – 30	30 – 40	40 – 50
Lince	15	45	32	22	18	60
Lima	50	32	28	35	44	22
Pueblo libre	15	36	45	32	60	18
Surco	40	24	14	46	45	24

a) Compare la variabilidad de las edades de los hombres y mujeres de Lince.

b) Compare la variabilidad de las edades en Lince y Pueblo Libre.

c) Compare la variabilidad de las edades de hombres y mujeres de la muestra.
d) Halle la varianza de las edades de toda la muestra.

Rp. a) CVH=0.187, CVM=0.212, b) CVL=0.235, CVPL=0.203, c) CVH=0.228, CVM=0.224, d) 62.072

33. Las notas finales del curso de Estadística que varían de cero a veinte, se tabularon en una distribución de frecuencias de 5 intervalos de igual amplitud. La ojiva de frecuencias relativas resultante corresponde al gráfico de la función:

$$H(x)=\begin{cases} 0.025x & 0\le x\le 4 \\ 0.0375x-0.05 & 4\le x\le 8 \\ 0.05x-0.15 & 8\le x\le 12 \\ 0.075x-0.45 & 12\le x\le 16 \\ 0.0625x-0.25 & 16\le x\le 20 \end{cases}$$

Calcule la media, la mediana y la desviación estándar de las notas

Rp. frec. 0.10, 0.15, 0.20, 0.30, 0.25, Media=11.8, Me=12.667,, Desv.estd=5.134.

34. Un gerente de ventas de una empresa ha registrado los siguientes montos de ventas de al menos S/.100 en un determinado día.

139, 222, 261, 209, 258, 204, 177, 227, 115.
154, 188, 233, 200, 247, 285, 241, 216, 220.
198, 181, 194, 102, 199, 215, 212, 209, 276.
218, 238, 197, 167, 223, 170, 194, 239, 205.
193, 267, 205, 199, 400, 300, 100, 102, 270.

Haga el análisis descriptivo de los datos utilizando diagrama de caja.

Rp. Min=100, Max=400, Media=210.4, s=54.6, Q_1=193, Me=209, Q_3=238 . RI=45, Min No Out=130. Max No Out=300, datos aislados: 400 (superior) y 100, 102 (inferior).

35. En una prueba de aptitud un investigador asigna a dos grupos de personas los siguientes "valores".

Grupo 1: 86, 89, 74, 73, 95, 86, 94, 90, 86, 88, 74, 76. 75, 93, 92
Grupo 2: 88, 95, 60, 100, 76, 93, 90, 98, 92, 75, 65, 68, 70, 73, 94.

Utilizando un diagrama de caja, compare la variabilidad y la asimetría de los dos grupos (¿Cuál de los dos es más asimétrica?)

a) Si los "valores" están en escala de intervalos
b) Si los "valores" están en escala de orden

Rp. a) Grupo1: Q_1=75, Me=86, Q_3=92 , RI=17, Min=73, Max=95, Grupo2: Q_1=70, Me=88, Q_3=94 . RI=24, Min=60, Max=100, Grupo2, es más disperso, ambos tienen asimetría negativa, En 1, 11/6=1.83 y en 2 18/6=3, luego 2 es más asimétrica, b) idem.

Capítulo 4

REGRESION LINEAL SIMPLE

4.1 Introducción

En este capítulo, trataremos con *muestras bivariantes* cuantitativas, es decir con muestras donde en cada unidad estadística se observan dos características cuantitativas medibles X e Y; por ejemplo, ingresos y gastos mensuales. El objetivo es estudiar la asociación entre dos variables conocida también como *asociación simple.*

La primera forma del estudio de la asociación entre las variables X e Y es la **regresión,** que consiste en determinar una relación funcional (recta de regresión) entre ellas, con el fin de que se pueda predecir el valor de una variable en base a la otra. La variable que se va predecir se denomina *variable dependiente* y la variable que es la base de la predicción se denomina *variable independiente.*

La segunda forma del estudio de la asociación entre las variables X e Y, es denominada **correlación**, que consiste en determinar la variación conjunta de las dos variables, su grado de relación, y su sentido (positivo o negativo). La medida del grado de relación se denomina *coeficiente o índice de correlación.* El cuadrado del índice de correlación se denomina **coeficiente de determinación.**

En este capitulo haremos un estudio descriptivo de las regresión lineal en el sentido que, la ecuación de regresión lineal que se determina será válida, si hay la seguridad de que existe un alto grado de correlación entre las variables indicado por el coeficiente de determinación.

Un estudio más avanzado de este tema se expone en el capitulo 13 del libro Estadística Inferencial: Aplicaciones, que viene a ser la segunda parte de este texto.

El lector debería correr paquetes de computo entre otros el *MCEST* para las aplicaciones de este capitulo.

4.1.1. Diagrama de dispersión

Sean $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ n valores de la variable bidimensional (X, Y), observados en una muestra, donde los x_i son los valores de la variable X y los y_i son los valores de la variable Y.

Los métodos estadísticos descriptivos son válidos en cada variable, es decir cada variable tiene media, desviación estándar, etc. Lo nuevo aquí es que con estos datos en pareja se puede medir la dispersión conjunta con respecto a las medias $(\bar{x}, \bar{y})$ mediante la covarianza.

Además, si los datos de X se tabulan en r intervalos; I_i; y los datos de Y se tabulan en s intervalos; I'_j, se tendrá una *distribución conjunta de frecuencias* que consiste de los intervalos (I_i, I'_j), y frecuencias f_{ij}. En este texto sólo haremos regresión con datos tabulados, pero no en intervalos.

Definición. Se denomina *diagrama de dispersión* o nube de puntos, a la gráfica de los valores (x_i, y_i) de las variables X e Y en el sistema cartesiano.

Es frecuentemente posible visualizar el tipo de relación existente entre dos variables a partir del diagrama de dispersión.

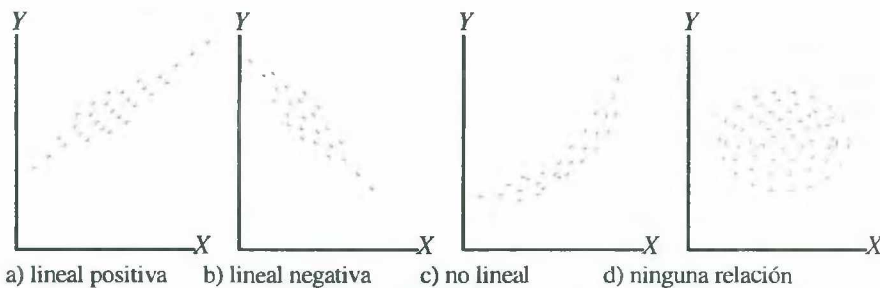


Fig. 4.1. Diagramas de dispersión

Por **ejemplo**, en la figuras 4.1 a),b) los datos visualizan una *relación lineal* entre las variables X e Y. En la figura 4.1 c) los datos visualizan una relación, pero, una *relación no lineal*, y en la figura 4.1 d) los datos visualizan ninguna relación válida en regresión entre las variables X e Y.

En este capítulo como ya se indicado en la introucción, haremos **regresión lineal descriptiva** determinando la ecuación lineal de regresión

$$Y = a + bX$$

que *mejor se ajusta* a los n pares de datos (x_i, y_i) y analizando la validez de la regresión a partir del **coeficiente de determinación**.

4.1.2 Covarianza

La covarianza es una estadística que mide el grado de dispersión o variabilidad conjunta de dos variables X e Y con respecto a sus medias respectivas $(\bar{x}, \bar{y})$.

Definición. La *covarianza* de n valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de una variable bidimensional (X, Y) es el número Cov(X,Y) o s_{XY} que se define igual a la media aritmética de los productos de las desviaciones de los datos con respecto a sus correspondientes medias $(\bar{x}, \bar{y})$. Esto es,

$$s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

En el numerador de s_{XY} se verifica la relación:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$$

Luego,
$$s_{XY} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{xy}.$$

La covarianza a diferencia de la varianza, **puede ser negativa**.

4.1.3 Coeficiente o índice de correlación

Definición. El coeficiente de correlación lineal de Pearson de n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de una variable bidimensional (X, Y). es el número abstracto r que se calcula por

$$r = \frac{s_{XY}}{s_X s_Y}$$

donde, s_{XY} es la covarianza de X e Y

s_X es la desviación estándar de X

s_Y es la desviación estándar de Y

El lector debería verificar que

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

Verificaremos (sección 4.2.3) que el coeficiente de correlación r es un número comprendido entre –1 y +1, esto es:

$$-1 \le r \le 1.$$

Interpretación:

Si $r = 1$, se dice que hay una correlación perfecta positiva.

Si $r = -1$, se dice que hay una correlación perfecta negativa.

Si $r = 0$, se dice que no hay correlación entre las dos variables.

4.2 Regresión lineal simple.

Dados n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de una variable bidimensional (X, Y). La regresión lineal simple de Y con respecto a X, consiste en determinar la ecuación de la recta:

$$Y = a + bX$$

que **mejor se ajuste** a los valores de la muestra, con el fin de poder *predecir* o *estimar* Y (variable dependiente) a partir de X (variable independiente).

El proceso de predecir o estimar Y a partir de la variable X, es la **regresión**. Hallar la función lineal $Y = a + bX$, consiste en determinar los valores de a y b a partir de los datos de la muestra.

Usaremos la notación $\hat{y}_i$ para representar un valor de Y calculado de la ecuación $Y = a + bX$ cuando X es igual a x_i. Esto es, $\hat{y}_i = a + bx_i$.

Al valor $\hat{y}_i$ se denomina *valor estimado* o *predecido* o *ajustado* de Y cuando $X = x_i$.

Si x_i es un valor de la muestra, entonces $(x_i, \hat{y}_i)$ es un punto de la recta de regresión $Y = a + bX$, (Fig. 4.2).

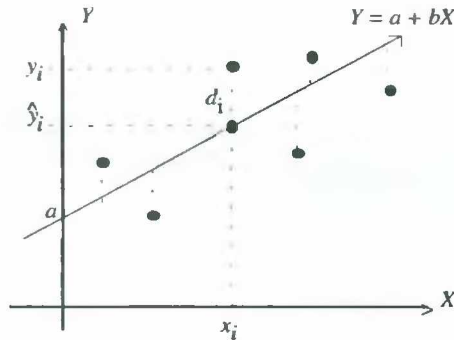


Fig. 4.2. Desviaciones de valores observados y ajustados

Definición. Se denomina **error o residuo** a cada diferencia,

$$d_i = y_i - \hat{y}_i$$

del valor observado y_i y el valor pronosticado $\hat{y}_i$ (Fig. 4.2).

Un método para determinar la recta que mejor se ajuste a los n datos de la muestra (x_i, y_i) es el método de *mínimos cuadrados*, que se explica a continuación.

4.2.1 Recta de regresión de mínimos cuadrados.

La recta de regresión de mínimos cuadrados de Y en X es aquella que hace mínima la suma de los cuadrados de errores (*SCE*) cuya expresión es:

$$SCE = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Luego, determinar una recta de regresión de mínimos cuadrados consiste en hallar los valores de a y b de manera que hagan mínima, la suma:

$$SCE = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

Este requisito se cumple, de acuerdo con el teorema de Gass-Markow, si a y b se determinan resolviendo el siguiente sistema de ***ecuaciones normales***:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

Estas ecuaciones **se obtienen de igualar a cero las derivadas de *SCE* con respecto a *a* y con respecto a *b*** respectivamente consideradas como variables, ya que (x_i, y_i) son datos observados.

Resolviendo el sistema de ecuaciones normales para b, se obtiene:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}, \quad \text{o} \quad b = \frac{s_{XY}}{s_X^2}$$

y dividiendo por n la primera ecuación normal, se tiene: el valor:

$$a = \bar{y} - b\bar{x}$$

NOTA. Sustituyendo $a = \bar{y} - b\bar{x}$ en $Y = a + bX$, resulta,

$$Y - \bar{y} = b(X - \bar{x})$$

que es otra forma de expresar la recta de regresión. Observar que la recta de regresión contiene al punto $(\bar{x}, \bar{y})$ cuyas componentes son las medias de X y de Y respectivamente.

Interpretación del coeficiente de regresión *b*

El coeficiente b es la *pendiente* o el *coeficiente de la regresión* lineal. La constante a es la ordenada en el origen.

Si $b > 0$, entonces, la tendencia lineal es creciente, es decir, a mayores valores de X corresponden mayores valores de Y. También, a menores valores de X corresponden menores valores de Y.

Si $b < 0$, entonces, la tendencia lineal es decreciente, es decir, a mayores valores de X corresponden menores valores de Y. También, a menores valores de X corresponden mayores valores de Y.

Si $b = 0$, entonces, $Y = a$. Luego, Y permanece estacionario para cualquier valor de X. En este caso se dice que, **no hay regresión**.

NOTA. *b* también se interpreta es el cambio promedio en $Y = a + bX$ cuando X cambia una unidad. Esto es, si x_i se incrementa 1, entonces $\hat{y}_i$ se incrementa en promedio b.

En general, si x_i se incrementa k, entonces $\hat{y}_i$ se incrementa en promedio kb (verificar!).

EJEMPLO 4.1.

En un estudio de la relación entre la publicidad por radio y las ventas de un producto, durante 10 semanas se han recopilado los tiempos de duración en minutos de la publicidad por semana (X), y el número de artículos vendidos (Y), resultando:

Semana	1	2	3	4	5	6	7	8	9	10
Publicidad X	20	30	30	40	50	60	60	60	70	80
Ventas Y	50	73	69	87	108	128	135	132	148	170

a) Trazar el diagrama de dispersión, e indicar la tendencia.
b) Calcular la recta de regresión de mínimos cuadrados con el fin de predecir las ventas.
c) Estimar la venta si en una semana se hacen 100 minutos de propaganda.
d) Calcular el coeficiente de correlación.
e) Si en la novena semana se incrementara la publicidad en 5 minutos, ¿en cuanto se estima se incrementen las ventas?.

SOLUCION.

a) Al trazar el diagrama de dispersión (fig. 4.3) vemos que hay una relación lineal positiva entre el número de artículos vendidos y el tiempo de publicidad semanal por radio.

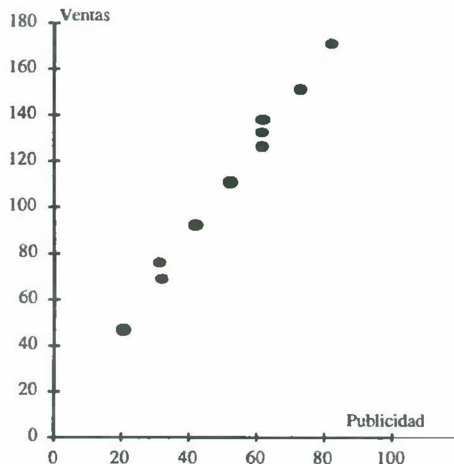


Fig. 4.3 Diagrama de dispersión

b) Para determinar la recta de regresión de mínimos cuadrados a partir de los datos. es decir para calcular a y b se dispone del cuadro 4.1.
De donde se obtiene:

$$n = 10,\ \sum X = 500,\ \sum Y = 1100,\ \sum XY = 61800$$

$$\sum X^2 = 28400, \qquad \sum Y^2 = 134660$$

$$\bar{x} = \frac{500}{10} = 50, \qquad \bar{y} = \frac{1100}{10} = 110$$

Cuadro 4.1. Computo de los coeficientes de regresión

X	Y	XY	X^2	Y^2
20	50	1000	400	2500
30	73	2190	900	5329
30	69	2070	900	4761
40	87	3480	1600	7569
50	108	5400	2500	11664
60	128	7680	3600	16384
60	135	8100	3600	18225
60	132	7920	3600	17424
70	148	10360	4900	21904
80	170	13600	6400	28900
500	1100	61800	28400	134660

Una forma de calcular b es:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{10 \times 61800 - 500 \times 1100}{10 \times 28400 - (500)^2} = \frac{68000}{34000} = 2$$

La otra forma del cálculo de b es:

$$s_{XY} = \frac{\sum XY}{n} - \overline{xy} = \frac{61800}{10} - 50 \times 110 = 680$$

$$s_X^2 = \frac{\sum X^2}{n} - \bar{x}^2 = \frac{28400}{10} - 50^2 = 340,$$

$$b = \frac{s_{XY}}{s_X^2} = \frac{680}{340} = 2.$$

Además, $$a = \bar{y} - b\bar{x} = 110 - 2(50) = 10.$$

La recta de regresión es: $Y = 10 + 2X$.

NOTA. Utilizando $Y - \bar{y} = b(X - \bar{x})$, se tiene:

$$Y - 110 = 2(X - 50) \quad \text{o} \quad Y = 10 + 2X$$

c) Si $x_i = 100$, $\hat{y} = 10 + 2(100) = 210$. No tenemos por el momento un criterio para concluir que este pronóstico es confiable.

d) El coeficiente de correlación

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{680}{18.439 \times 36.959} = 0.9978$$

es altamente positivo. Es un primer criterio para analizar la validez de la predicción

e) Si en la novena semana se incrementara el tiempo de propaganda en 5 minutos, entonces, la venta se incrementa en promedio 5×2=10 unidades..

EJEMPLO 4.2.

Los ingresos (X) y los gastos (Y) mensuales en dólares de una muestra de 100 familias han dado los siguientes resultados

$$\bar{x} = 210,\ \bar{y} = 200,\ s_X^2 = 5.76,\ s_Y^2 = 2.56,\ \sum XY = 4200364.8$$

Determine la recta de regresión de mínimos cuadrados de Y en X y estime el gasto de una familia que tiene $250 de ingreso.

SOLUCION.

a) $$s_{XY} = \frac{\sum XY}{n} - \bar{x}\bar{y} = \frac{4200364.8}{100} - 210 \times 200 = 3.648$$

La recta de regresión de mínimos cuadrados de Y en X es

$$Y - \bar{y} = \frac{s_{XY}}{s_X^2}(X - \bar{x}),$$

$$Y - 200 = \frac{3.648}{5.76}(X - 210),$$

$$Y = 67 + 0.633X$$

b) Si una familia tiene un ingreso de $250 entonces su gasto estimado sería

$$\hat{y} = 67 + 0.633 \times 250 = 225.25.$$

NOTAS:

1) De $b = s_{XY}/s_X^2$ y $r = s_{XY}/s_X s_Y$, se obtiene la relación entre los coeficientes de correlación r, y el de regresión; b;

$$b = r\frac{s_Y}{s_X}$$

Entre otras cosas, r y b *tienen el mismo signo.*

2) La recta de regresión de X en Y, es decir X variable dependiente de Y está dada por:

$$X = c + dY$$

donde, $$d = \frac{s_{XY}}{s_Y^2} \quad \text{y} \quad c = \bar{x} - d\bar{y}$$

Esta recta de regresión de X en Y se puede escribir también como:

$$X - \bar{x} = d(Y - \bar{y}) \text{ ó } Y - \bar{y} = \frac{1}{d}(X - \bar{x})$$

Observar que también pasa por el punto $(\bar{x}, \bar{y})$

3) Los coeficientes de regresión b y d verifican:

$$bd = \frac{s_{XY}}{s_X^2}\frac{s_{XY}}{s_Y^2} = \frac{s_{XY}}{s_X s_Y}\frac{s_{XY}}{s_X s_Y} = rr = r^2$$

El número r^2 es denominado **coeficiente de determinación.**

4) Comparando las rectas de regresión:

$$L_1 : Y - \bar{y} = b(X - \bar{x}) \text{ de } Y \text{ en } X$$

$$L_2 : Y - \bar{y} = \frac{1}{d}(X - \bar{x}) \text{ de } X \text{ en } Y$$

resulta que, son *coincidentes* si $bd = 1$, o si $r^2 = 1$.

Por otra .parte, $r = 0$, significa que L_1 (o L_2) es paralela al eje X, L_2 (o L_1) es paralela al eje Y y perpendiculares entre si en el punto común $(\bar{x}, \bar{y})$. En consecuencia, si r tiende a cero, las rectas L_1 y L_2 tienden a ser perpendiculares y si r tiende a 1 o a -1, las rectas L_1 y L_2 tienden a ser coincidentes.

4.2.2 Partición de la varianza de Y, s_Y^2

Sea (x_i, y_i) un valor observado de la variable (X, Y) e $\hat{y}_i$ el valor en la ecuación de regresión $Y = a + bX$ cuando $X = x_i$

La varianza de Y es el número:

$$s_Y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}.$$

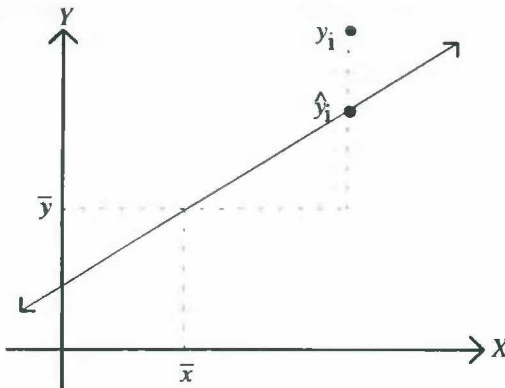


Fig. 4.4

Observar que en la figura 4.4 se tiene:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Error total =Error no Explicado + Error explicado por la regresión.

Esta terminología surge, debido a que las desviaciones $y_i - \hat{y}_i$ con respecto a la recta de regresión, se comportan de una manera aleatoria o impredecible, debido

a que y_i es aleatorio. En tanto que las desviaciones $\hat{y}_i - \overline{y}$ de la recta de regresión con respecto al eje de las X se explican por la recta de regresión de Y en X, ya que sólo depende de los $\hat{y}_i$ que están sobre la recta.

Por otro lado se verifica la siguiente partición de sumas de cuadrados:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

$$SCT = SCE + SCR$$

En esta partición de las sumas, la primera suma se denomina ***suma de cuadrados total*** (*SCT*), refleja la variación de los valores de Y con respecto a la media $\overline{y}$.

La segunda suma se denomina, ***suma de cuadrados de los errores*** (*SCE*), y la tercera suma se denomina ***suma de cuadrados debido a la regresión*** (*SCR*), refleja la cantidad de variación de los valores de *Y explicada* por la recta de regresión.

Si se divide por *n*, (el tamaño de la muestra), entonces, se dice que la "*varianza de los y_i es igual a la varianza no explicada o residual más la varianza explicada por la recta de regresión*".

EJEMPLO 4.3.

En una muestra de 5 obreros de una fábrica se han observado sus años de experiencia (X) y el tiempo que tardan en realizar una determinada tarea (Y). Los datos se muestran en la tabla que sigue:

X	1	2	3	4	5
Y	8	9	4	3	3

Verificar que la variación total es igual a la variación no explicada más la variación explicada por la regresión de Y en X.

SOLUCION.

De los datos de la muestra se obtiene la siguiente ecuación lineal de regresión por mínimos cuadrados.

$$Y = 10.2 - 1.6X \qquad \text{donde} \quad \overline{y} = 5.4$$

Cuadro 4.2

y_i	$\hat{y}_i$	$y_i - \overline{y}$	$y_i - \hat{y}_i$	$\hat{y}_i - \overline{y}$	$(y_i - \overline{y})^2$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \overline{y})^2$
8	8.6	+2.6	−0.6	+3.2	6.76	0.36	10.24
9	7.0	+3.6	+2.0	+1.6	12.96	4.00	2.56
4	5.4	−1.4	−1.4	0.0	1.96	1.96	0.00
3	3.8	−2.4	−0.8	−1.6	5.76	0.64	2.56
3	2.2	−2.4	+0.8	−3.2	5.76	0.64	10.24
27		0.0	0.0	0.0	33.20	7.60	25.60

Del cuadro 4.2, resulta

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$33.20 = 7.60 \quad + \quad 25.60.$$

NOTA. Para comparar estas varianzas se convierten a **varianza relativas**, dividiendo la identidad entre 33.20 (SCT).

$$\frac{33.20}{33.20} = \frac{7.60}{33.20} + \frac{25.60}{33.20}$$

$$1 = 0.23 + 0.77$$

La lectura es como sigue:

El 100% de la varianza total se particiona en 23% de varianza no explicada más 77% de varianza explicada por la regresión de *Y* en *X*.

4.2.3 Coeficiente de determinación

El coeficiente de determinación r^2 se define como el cociente: $\frac{SCR}{SCT}$.

Esto es,. el *coeficiente de determinación* r^2 de la regresión de Y en X, está dada por la expresión:

$$r^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

De la partición de suma de cuadrados, $SCT = SCE + SCR$, resulta:

$$1 = \frac{SCE}{SCT} + r^2 \qquad (*)$$

Por lo tanto para interpretar la partición de varianzas relativas bastará con calcular r, luego, r^2 y establecer:

$$1 = (1 - r^2) + r^2$$

para concluir que **el 100% de la varianza total es igual** $(1-r^2)\times 100\%$ **de varianza no explicada más** $r^2 \times 100\%$ **de la variación explicada por la recta de regresión.**

Por **ejemplo**, en el ejemplo 4.2, $r=0.95$, r^2=0.9025, entonces, se tiene:

$$1=(1-r^2)+r^2$$

$$1=0.0975+0.9025 \text{ o aproximando a dos decimales } 1=0.10+0.90$$

Es decir, el 90% de la variabilidad en los gastos mensuales se explica por la asociación con los ingresos mensuales. Quedan 10% de variabilidad en los gastos que no se explica por la regresión.

Consecuencias.

1) De la identidad (*) se concluye que $0 \le r^2 \le 1$. Entonces, $-1 \le r \le 1$.
   Si $r>0$, se dice que existe una correlación *directa positiva*, ambas variables aumentan (o disminuyen) simultáneamente.
   Si $r<0$, se dice que existe una correlación *inversa negativa*, mientras los valores de una variable aumenta, los de la otra disminuyen y viceversa.
   Si $r=0$, se dice que no hay correlación entre X e Y. Por lo tanto no hay regresión de Y en X.

2) $r^2=1$, sólo si, $SCE=0$, o sólo si, $y_i=\hat{y}_i$ para los n datos de la muestra. Esto significa que todos los y_i están en la recta de regresión. En este caso se dice que hay una correlación perfecta entre X e Y.
   Si $r=1$, se dice que hay una *correlación perfecta positiva.*
   Si $r=-1$, se dice que hay una *correlación perfecta negativa.*

3) $r^2=0$, sólo si, $SCR=0$, o sólo si, $\hat{y}_i=\bar{y}$ para los n datos de la muestra. Es decir y_i no cambia cuando cambia x_i, o todas las predicciones son iguales a una misma constante. En este caso **no hay correlación ni regresión.**

4) El coeficiente de determinación r^2, es pues una medida de la proximidad del ajuste de la recta de regresión. *Cuanto mayor sea el valor de* r^2, *mejor será el ajuste y más útil la recta de regresión como instrumento de predicción.* ($r^2=0.90$ indica que de 100 pares de puntos 90 están en la recta de regresión y 10 fuera de la recta de regresión)

NOTA. (Una advertencia)

El haber supuesto una función lineal entre dos variables y haber encontrado un alto coeficiente de correlación, no necesariamente significa que una variable

dependa de la otra, pues, esta correlación puede no ser *causal* si no *casual* . Para que exista correlación debe haber causa y efecto.

4.2.4 Invarianza del coeficiente de regresión y del indice de correlación

Hay dos formas de simplificar los cálculos de la pendiente b de la regresión y del coeficiente de correlación r. La primera forma es mediante un *cambio de origen*, que consiste en transformar los datos de X o Y o de ambos a la forma $X' = X - h$, $Y' = Y - k$ con h y k constantes.

Si se hacen tales transformaciones en X o Y o en ambos, b y r no cambian, esto es:

$$b = \frac{s_{XY}}{s_X^2} = \frac{s_{X'Y'}}{s_{X'}^2} = \frac{s_{X'Y}}{s_{X'}^2} = \frac{s_{XY'}}{s_X^2}$$

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{s_{X'Y'}}{s_{X'} \cdot s_{Y'}} = \frac{s_{X'Y}}{s_{X'} \cdot s_Y} = \frac{s_{XY'}}{s_X s_{Y'}}.$$

La segunda forma de simplificar el cálculo de b y de r, es mediante una *reducción de la escala*, que consiste en dividir X e Y por una constante diferente de cero.

Si $X' = X/h$ e $Y' = Y/h$, entonces, la pendiente b no cambia, esto es,

$$b = \frac{s_{XY}}{s_X^2} = \frac{s_{X'Y'}}{s_{X'}^2}$$

Si se hace la transformación $X' = X/h$, $Y' = Y/k$, en X o Y o en ambos entonces, el coeficiente r no cambia, esto es,

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{s_{X'Y'}}{s_{X'} \cdot s_{Y'}} = \frac{s_{X'Y}}{s_{X'} \cdot s_Y} = \frac{s_{XY'}}{s_X s_{Y'}}.$$

EJEMPLO 4.4

El ingreso anual disponible y los gastos de consumo (en dólares) de una muestra de 10 familias de un barrio residencial de Lima fueron tabulados en el cuadro que sigue.

Hallar la recta de regresión del consumo (Y) con respecto al ingreso (X), utilizando la transformación $X' = X/1{,}000$, $Y' = Y/1{,}000$.

Ingreso	Consumo
20,000	18,000
14,000	15,000
35,000	30,000
23,000	16,000
12,000	9,000
5,000	7,000
7,000	7,000
14,000	15,000
30,000	26,000
25,000	23,000

SOLUCION.

Realizamos la codificación: $X' = X/1{,}000$, $\quad Y' = Y/1{,}000$, del cuadro 4.3 se obtiene:

Cuadro 4.3 . Cálculos para la regresión lineal con datos codificados

X'	Y'	$X'Y'$	X'^2	Y'^2
20	18	360	400	324
14	15	210	196	225
35	30	1050	1225	900
23	16	368	529	256
12	9	108	144	81
5	7	35	25	49
7	7	49	49	49
14	15	210	196	225
30	26	780	900	676
25	23	575	625	529
185	166	3745	4289	3314

$$\sum X' = 185,\ \sum Y' = 166,\ \sum X'Y' = 3745,\ \sum X'^2 = 4289,\ \sum Y'^2 = 3314$$

$$\bar{x}' = 18.5,\quad \bar{y}' = 16.6$$

$$b = \frac{n\sum X'Y - \sum X'\sum Y}{n\sum X'^2 - (\sum X')^2} = \frac{10(3745) - 185(166)}{10(4289) - (185)^2} = 0.778$$

La recta de regresión de Y en X es:

$$Y - \bar{y} = b(X - \bar{x}),$$

donde $\quad b = 0.778,\quad \bar{y} = 1000 \times \bar{y}' = 16{,}600,\quad \bar{x} = 1000 \times \bar{x}' = 18{,}500$

Resultando, $\quad Y = 2{,}207 + 0.778X$.

EJEMPLO 4.5

Al estudiar la relación entre costos (X) y ventas(Y) en dólares de ciertos productos, a partir de una muestra se obtuvo la siguiente información:

$$s_X = 5,\ s_Y = 4,\ \bar{x} = 50,\ \bar{y} = 100,\ Y = 62 + 0.76X$$

Si los costos se incrementan en \$3 y las ventas correspondientes se incrementan en 6 \$

a) ¿Cómo cambia la ecuación de regresión?.

b) ¿Qué porcentaje de la varianza de las ventas es explicada por la regresión de ventas sobre costos?.

SOLUCION

a) Si $X' = X + 3, Y' = Y + 6$, la ecuación de regresión de Y' en X' es:

$$Y' - \bar{y}' = b'(X' - \bar{x}')$$

donde $b' = b = 0.76$, $\bar{y}' = \bar{y} + 6 = 106$, $\bar{x}' = \bar{x} + 3 = 53$. Esto es

$$Y' - 106 = 0.76(X' - 53)$$

b) Se tiene: $b = 0.76$, de $r = \dfrac{s_{XY}}{s_X s_Y}$ y $b = \dfrac{s_{XY}}{s_X^2}$ resulta $r = 0.95$.

Si se hace la transformación $X' = X + 3, Y' = Y + 6$, el coeficiente de correlación r=0.95 no cambia. Por lo tanto, el porcentaje de la varianza de Y (o de Y') explicada por la regresión de Y en X (o de Y' en X') es la misma:

$$r^2 = (0.95)^2 = 0.9025.$$

EJEMPLO 4.6 (Aplicación a serie de tiempo)

Cuando una de las variables es el tiempo (en días, meses o años), la regresión se denomina *serie de tiempo*.

Supongamos que la producción (en millones) de un determinado artículo fabricado por una compañía durante los años 1980-1989 es como sigue:

Años	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Producción	92.2	92.3	80.0	89.1	83.5	68.9	69.2	67.1	58.3	61.2

a) Trazar un gráfico de líneas y describir la tendencia.

b) Hallar la recta de regresión (serie de tiempo) de mínimos cuadrados de la producción en función de los años.

c) Estimar la producción de artículos para 1990 y establecer si es significativa tal predicción.

SOLUCION.

a)La figura 4.5 es la gráfica de líneas de la producción por años.

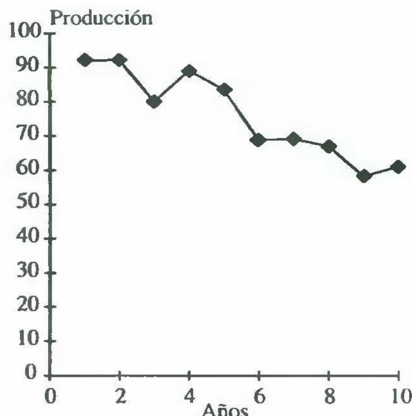


Fig. 4.5 Producción de los artículos de 1980 a 1989

b) Haciendo $X' = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,$ para los años respectivos:
1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989,
se tiene:

$$\sum X' = 55,\ \sum Y = 761.8,\ \sum X'Y = 3862,\ \sum X'^2 = 385,\ \sum Y^2 = 59513.785$$

$$b = \frac{n\sum X'Y - \sum X'\sum Y}{n\sum X'^2 - (\sum X')^2} = \frac{10(3862) - 55(761.8)}{10(385) - (55)^2} = -3.97$$

$$a = \bar{y} - b\bar{x}' = 76.18 - (-3.97)(5.5) = 98.015$$

La recta de regresión es: $Y = 98 - 3.97X'$, con $r = 0.94$.

c) Para el año 1990, $x_i' = 11$, la producción para ese año será:

$$\hat{y}_i = 98 - 3.97x_i' = 98 - 3.97 \times 11 = 54.33.$$

4.3 Nociones de regresión no lineal

En muchos casos cuando los valores en parejas de las variables X, e Y, no se ajustan a una línea recta, se puede conseguir una relación lineal mediante una transformación de estos valores.

A continuación se dan algunas ecuaciones no lineales y su transformación lineal

Ecuación **Transformación lineal**

a) $Y = AB^X$ (exponencial) $\log Y = \log A + (\log B)X$

b) $Y = AX^B$ (potencia) $\log Y = \log A + B \log X$

c) $Y = 1/(A + BX)$ (hiperbólica) $Y' = A + BX$, siendo $Y' = 1/Y$

EJEMPLO 4.7

Ajustar por el método de mínimos cuadrados una curva de la forma

$$Y = AX^B$$

a los siguientes pares de datos:

X	1.5	2	3	3.5	4	5
Y	2.6	2.4	1.2	1.8	1.6	1.4

SOLUCION.

La transformación a la regresión lineal es:

$$Y' = A' + B'X'$$

donde: $Y' = \log Y$, $X' = \log X$, $A' = \log A$, $B' = B$

De la tabla 4.4 se obtiene:

$\sum X' = 2.7993$, $\sum Y' = 1.4799$, $\sum X'Y' = 0.5891$, $\sum X'^2 = 1.4962$, $\sum Y'^2 = 0.4513$

Cuadro 4.4 . Cálculos para la regresión no lineal

X	Y	X'=log X	Y'=log Y	$X'Y'$	X'^2	Y'^2
1.5	2.6	0.1761	0.4150	0.0731	0.0310	0.0.172
2.0	2.4	0.3010	0.3802	0.1144	0.0906	0.1446
3.0	1.2	0.4771	0.0792	0.0378	0.2276	0.0063
3.5	1.8	0.5440	0.2553	0.1378	0.2959	0.0652
4.0	1.6	0.6021	0.2041	0.1229	0.3625	0.0417
5.0	1.4	0.6990	0.1461	0.1021	0.4886	0.0213
		2.7993	1.4799	0.5891	1.4962	0.4513

$$b = \frac{n\sum X'Y' - \sum X'\sum Y'}{n\sum X'^2 - (\sum X')^2} = \frac{6(0.5891) - (2.7993)(1.4799)}{6(1.4962) - (2.6993)^2} = -0.532$$

$$a = \overline{y}' - b\overline{x}' = 0.24665 - (-0.532)(0.46655) = 0.4949$$

Además, $$r = \frac{s_{X'Y'}}{s_{X'} \cdot s_{Y'}} = \frac{-0.608}{(1.068)(0.7196)} = -0.79.$$

La ecuación lineal de regresión es:

$$Y' = 0.4949 - 0.532X'$$

La ecuación no lineal de regresión se obtiene utilizando antilogaritmos:

$$Y = anti\log(Y') = (3.125)X^{-0.532}$$

$$\text{donde} \quad A = anti\log(a) = anti\log(0.4949) = 3.125 .$$

EJEMPLO 4.8.

Para los siguientes datos experimentales

X	1	2	3	4	5	6
Y	10	40	120	300	800	1500

Se plantean los modelos:

$$Y = Ae^{BX} \quad \text{e} \quad Y = a + bX ,$$

para relacionar Y con X, ¿cuál de los dos modelos se ajusta mejor a los datos?

SOLUCION.

Si ajustamos a los datos la ecuación no lineal $Y = Ae^{BX}$, su transformación lineal es $\ln Y = \ln A + BX$, esto es

$$Y' = a + BX, \quad \text{donde} \quad Y' = \ln Y, \quad a = \ln A .$$

De los datos experimentales se obtiene:

$$\sum X = 21, \quad \sum Y' = 30.481, \quad \sum XY' = 124.16, \quad \sum X^2 = 91, \quad \sum Y'^2 = 172.53$$

$$B = \frac{n\sum XY' - \sum X \sum Y'}{n\sum X^2 - [\sum X]^2} = 0.99876,$$

$$a = \overline{y}' - B\overline{x} = 1.58443 .$$

$$r = 0.9935$$

La ecuación lineal de regresión es: $Y' = 1.58443 + 0.99876X$

La ecuación no lineal de regresión es: $Y = anti\ln(Y') = (4.8765)e^{(0.99876)X}$

Si ajustamos la ecuación lineal $Y = a + bX$, de los datos se obtiene:

$$\sum X = 21, \quad \sum Y = 2770, \quad \sum XY = 14650, \quad \sum X^2 = 91, \quad \sum Y^2 = 2996100$$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - [\sum X]^2} = 283.143, \quad a = \overline{y} - b\overline{x} = -529.33, \quad r = 0.904$$

La ecuación lineal de regresión es: $Y = -529.33 + 283.143X$

Finalmente comparando los coeficientes de determinación: $r^2 = 0.987$, para el modelo no lineal y $r^2 = 0.817$ para el modelo lineal, se concluye que el modelo no lineal se ajusta mejor.

EJERCICIO 1.

Si $d_i = y_i - \hat{y}_i = y_i - a - bx_i$, verificar que:

a) $\sum_{i=1}^{n} d_i = 0$, b) $\sum_{i=1}^{n} d_i x_i = 0$, c) $\sum_{i=1}^{n} d_i \hat{y}_i = 0$, d) $\sum_{i=1}^{n} (\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 0$

SOLUCION.

a) $\sum d_i = \sum (y_i - \hat{y}_i) = \sum (y_i - a - bx_i) = \sum y_i - na - b\sum x_i = 0$

La última identidad es igual a cero por ser la primera ecuación normal.

b) $\sum d_i x_i = \sum (y_i - a - bx_i)x_i = \sum x_i y_i - a\sum x_i - b\sum x_i y_i = 0$

La última identidad es igual a cero por ser la segunda ecuación normal.
Use a) y b) para probar c). Use a), b) y c) para probar d).

EJERCICIO 2.

Verificar que

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

SOLUCION

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2$$

$$\sum (y - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

Por otro lado, $2\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$, por d) del ejercicio 1.

EJERCICIO 3.

Demostrar que si $r = \frac{s_{XY}}{s_X s_Y}$, entonces, $r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$.

SOLUCION.

$$r^2 = \frac{s_{XY}}{s_X s_Y}\frac{s_{XY}}{s_X s_Y} = \frac{s_{XY}}{s_X^2}\frac{s_{XY}}{s_Y^2} = b\frac{s_{XY}}{s_Y^2} = \frac{b\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}$$

Se debe probar que

$$b\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(\hat{y}_i - \bar{y})^2$$

En efecto,

$$\sum(y_i - \hat{y}_i)^2 = \sum(y_i - a - bx_i)^2 = \sum(y_i - \bar{y} + b\bar{x} - bx_i)^2$$
$$= \sum[(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$
$$= \sum(y_i - \bar{y})^2 + b^2\sum(x_i - \bar{x})^2 - 2b\sum(x_i - \bar{x})(y_i - \bar{y})$$

Remplazando $b = s_{XY}/s_X^2$ la segunda y tercera suma del segundo lado de la igualdad se reducen a $-b\sum(x_i - \bar{x})(y_i - \bar{y})$. Luego,

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y})^2 + b\sum(x_i - \bar{x})(y_i - \bar{y})$$

Por otro lado, se tiene

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$$

Comparando las 2 últimas identidades se tiene el resultado deseado

EJERCICIOS

1. ¿ Por qué son iguales los signos del coeficiente de correlación y de la pendiente de la recta de regresión de mínimos cuadrados?. Rp. $r=bs_X/s_Y$.

2. Dadas las rectas de regresión de mínimos cuadrados $Y = a + bX$ y $X = c + dY$. Verificar que bd es igual al coeficiente de determinación.

3. Si las gráficas de las rectas de regresión de Y en X y de X en Y forman un ángulo de 90 grados, ¿qué se puede afirmar del índice de correlación?. Rp. es cero.

4. Si $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ son n pares de datos observados que se encuentran en la recta L: $Y = mX + b$, ¿por qué L es la recta de regresión de mínimos cuadrados para estos puntos?, ¿qué porcentaje de la varianza total de los y_i es explicado por L?.

   Rp. por que $\Sigma(y - \hat{y}_i)^2=0$, 100%.

5. Dada la recta de regresión de mínimos cuadrados $Y = a + bX$, si se produce un incremento igual a c en uno de los valores de X, ¿cuánto es el incremento respectivo que se produce en Y?. Rp. bc.

6. Al realizar la regresión de Y en X basado en una muestra de 10 pares de datos (x_i, y_i), se tiene que la varianza de los y_i es igual a 16 y que la suma de cuadrados debido a la regresión es 140. ¿Qué porcentaje de la varianza de los y_i es explicada por la regresión ?. Rp. 87.5%.

7. El coeficiente de correlación entre dos variables X e Y es $r = 0.60$. Si $s_X = 1.50$, $s_Y = 2.00$, $\bar{x} = 10$, $\bar{y} = 20$, hallar la recta de regresión:
   a) de Y en X, b) de X en Y.

   Rp. a) $Y = 12 + 0.8X$, b) $X = 1 + 0.45Y$, o $Y = -2.2 + 2.2X$

8. Si n pares $(x_1, y_1), \ldots, (x_n, y_n)$ son tales que cumplen la relación $y_i = \frac{1}{5}(x_i - 2)$, y si $\bar{y} = 0$, ¿es válida la relación $S_{XY} = S_X^2 / 5$?. Rp. si

9. Si la ecuación de regresión de Y en X es : $Y = 3 + 2X$, hallar la ecuación de regresión de Y' en X', donde $X' = X + 3$ e $Y' = Y + 6$.

   Rp. $Y'-9 = 2(X' - 3)$

10. Al estudiar la regresión lineal entre los ingresos medios (Y en \$) y el número de hijos por familia (X), se obtuvo la siguiente información:

$$\bar{x}=3,\ \ \bar{y}=700,\ s_X=0.5\sqrt{s_{XY}}\ ,$$

estimar los ingresos de las familias con 4 hijos, ¿a cuántos hijos por familia correspondería un ingreso estimado en \$712?.

Rp. $Y=688+4X$, $\hat{y}=\$704$, $x=6$

11. Al estimar las ventas (Y) de un artículo en función de los precios (X) se usó una recta de mínimos cuadrados basado en una muestra de 4 datos. Si las ventas observadas fueron 10, 8, 6, 14 y si las ventas estimadas respectivas son 10.8, 8.2, 5.6, 13.4, ¿qué porcentaje de la varianza de las ventas es explicada por la recta de regresión?. Rp. 96.57%.

12. En un estudio de la relación entre ingresos mensuales y gastos de educación de las familias, una muestra proporciona un coeficiente de determinación del 90.25%, medias respectivas de \$420 y \$120, y desviaciones estándar respectivas de \$10 y 7\$. Según este estudio

a) ¿En cuánto se estima los gastos por educación de una familia cuyo ingreso mensual es de \$300?.

b) Si una familia estima su gasto por educación en \$370, ¿cuánto debería ser su ingreso mensual?.

c) Si una familia tiene un aumento de \$50, ¿en cuanto se incrementaría la estimación de sus gastos en educación?.

Rp. $Y=-159.3+0.6665X$, a) \$40.2, b) \$795.94, $0.665\times 50=33.25$.

13. Al estudiar la relación entre la edad (X) y la presión sanguínea (Y) a partir de una muestra de mujeres, se obtuvo la siguiente información:

$$s_X=7.5,\ s_Y=10,\ \bar{x}=50,\ \bar{y}=120,\ r=0.90$$

a) Hallar la relación lineal de la presión con respecto a la edad y predecir la presión sanguínea para una mujer de 45 años.

b) ¿Qué porcentaje de la varianza de la presión es explicada por la regresión de la presión sanguínea con respecto a la edad?.

Rp. a) $Y=60+1.2X$, si $x=45$, $\hat{y}=114$, b) 81%.

14. Al estudiar la relación entre el costos(X) y las utilidades (Y) en dólares de ciertos productos a partir de una muestra se obtuvo la siguiente información:

$$s_X=5,\ s_Y=4,\ \bar{x}=100,\ \bar{y}=50,\ Y=-26+0.76X.$$

a) ¿Qué porcentaje de la varianza de las utilidades es explicada por la regresión de utilidades sobre costos?.

b) Si cada valor del costo se aumenta en \$3 y el valor correspondiente a la utilidad se aumenta en 6 \$, ¿en cuanto se estima la utilidad para un costo de \$120?.

Rp. a) r=0.95, r^2=0.9025, el 90.25%, b) $Y'-56=0.76(X'-103)$, 68.92.

15. Utilizando los n pares de datos $(x_1, y_1),\ldots,(x_n, y_n)$ se obtiene una ecuación de regresión lineal con pendiente igual a -1. Determinar el índice de correlación entre los valores x_i e y_i si además, $s_X / s_Y = 0.9$.

Rp. $r = -0.9$.

16. Si n pares $(x_1, y_1),\ldots,(x_n, y_n)$ tienen índice de correlación r, comprobar que la recta de regresión para los puntos $(x_1^*, y_1^*),\ldots,(x_n^*, y_n^*)$, en donde $x_i^* = \dfrac{x_i - \overline{x}}{s_X}$, $y_i^* = \dfrac{y_i - \overline{y}}{s_Y}$ para $i = 1,2,\ldots,n$, es $y^* = rx^*$.

Rp. $\overline{x^*} = 0$, $\overline{y^*} = 0$, $r^* = r$, entonces $y^* = rx^*$

17. Si n pares $(x_1, y_1),\ldots,(x_n, y_n)$ son tales que cumplen la relación: $Y = bX$, deducir b usando el método de mínimos cuadrados. Rp. $b=\sum xy/\sum x^2$

18. Una compañía de alimentos maneja una cadena de tiendas al menudeo. Para medir la eficiencia de las tiendas se estudió la relación del número de empleados (X) y el promedio del volumen de ventas mensuales (Y) expresado en cientos de dólares para todas las tiendas durante el año pasado. La gráfica de los datos sugiere una relación lineal entre las variables. Se tiene la siguiente información:

$$n = 100,\ \sum X = 600,\ \sum Y = 1600,\ \sum XY = 13600,\ \sum X^2 = 5200,\ \sum Y^2 = 37700$$

a) Hallar la recta de mínimos cuadrados para estimar las ventas partir del número de empleados. ¿En cuánto se estiman las ventas para una tienda de 8 empleados?.

b) ¿Qué porcentaje de la varianza de las ventas es explicada por la variabilidad del número de empleados?.

c) ¿Cuántos empleados tiene la tienda cuya venta promedio se estima en \$1,100?.

Rp. a) Y=1+2.5X, 21 o \$2100, b) r=0.909 , 82.6%, c) 4.

19. Un estudio de mercado trata de averiguar si es efectiva la propaganda televisada de un producto que salió a la venta con relación al tiempo de publicidad (en horas/semana). Se recopilaron datos a partir de la segunda semana de iniciada la publicidad resultando el cuadro que sigue. No se pudo recopilar datos de la cuarta semana.

Semana	2	3	4	5	6	7
Tiempo de propaganda	20	25	22	28	36	40
Venta del Producto($)	300	310	–	320	350	420

a) ¿Es efectiva la publicidad del producto?.
b) ¿En cuanto estimaría las ventas para la semana 4?.

Rp. a) $Y = 176.82 + 5.467X$, $r = 0.92$, b) 297.

20. Un editor tomó una muestra de 7 libros anotando el precio y el número de páginas respectivo, obteniendo los siguientes datos:

No. de páginas	630	550	400	250	370	320	610
Precio ($)	10	8	7	4	6	6	9

a) Determine una función lineal entre el precio y el número de páginas con el fin de predecir precios. ¿ Que porcentaje de la varianza total de precios se explica por esta función?.
b) Estimar el precio de un libro de 300 páginas. Si a este libro se le incrementa 20 páginas en una segunda edición, ¿en cuánto se incrementaría su precio?.
c) ¿Cuántas páginas debería tener un libro cuyo precio se estima en $12.27?.

Rp. a) $Y = 1.22 + 0.013X$; 94.5%, b) $5.12, $ 0.26, c) 850.

21. Una muestra de 5 varones adultos de quienes se observaron las estaturas (X en pies, pulgadas) y los pesos (Y en libras) ha dado los siguientes resultados:

X	5'1"	5'2"	5'3"	5'4"	5'5"
Y	125	130	140	145	160

a) Realice una regresión lineal y utilice los datos para verificar que la varianza total de Y es igual a la varianza residual más la varianza explicada por la recta de regresión.
b) Usando la descomposición de la varianza calcule r^2 e interprete el resultado.

Rp. a) $Y = 114.5 + 8.5X'$, $X' = 1,2,3,4,5$. $750 = 27.5 + 722.5$, b) $r^2 = 0.96$.

22. Se quiere estudiar la relación entre las edades en años (X) de un tipo de máquinas que se utiliza en la fabricación de cierto artículo y el número de artículos (Y) que producen. A partir de la muestra de la tabla siguiente:
a) Determinar la recta de regresión de mínimos cuadrados para predecir la producción. Estimar la producción para 4, 7 y 8 años.
b) Calcular el porcentaje de la varianza explicada por la regresión de la producción.
c) Si realmente cada máquina de la muestra produce 10 artículos menos determinar la recta de regresión. ¿Cuánto es el porcentaje de la varianza explicada por la regresión de la producción?.

X	Y
2	95
3	70, 80
4	–
5	75
6	60
7	–
8	–
9	45, 50
10	25

Rp. a) $Y = 101.5 - 6.64X$, b) 88%, c) $Y' = 91.5 - 6.64X$ y r no varía.

23. Sea Y el índice de precios al consumidor, tomando como base al año 1980 (es decir 1980 = 100). Para los datos que siguen:

Año	1981	1982	1983	1984	1985	1986	1987
Y	106.0	111.1	117.2	121.3	125.2	128.0	132.6

a) Hallar la recta de mínimos cuadrados que se ajuste a los datos.
b) Predecir el índice de precios para el año 1988 y compararlo con el valor verdadero (144.4). ¿En que año podemos esperar que el índice de precios sea 150.57, suponiendo que las tendencias presentes continúen?.

Rp. a) $Y = 102.83 + 4.34X$, $r = 0.994$ b) año 1991.

24. Los porcentajes en gastos de publicidad y los porcentajes de beneficios netos de ventas en una muestra de 9 negocios es como sigue:

Gastos	2.3	1.9	3.5	1.0	1.5	4.0	2.6	3.0	2.4
Beneficios	4.0	3.8	6.2	2.9	3.4	6.8	4.5	5.0	4.2

a) Hallar la recta de regresión de mínimos cuadrados para predecir beneficios netos.
b) Determine el beneficio si el gasto es 5%. ¿Cuánto es el porcentaje de la varianza explicada de los beneficios con respecto al gasto?.

Rp. a) $Y = 1.274 + 1.32X$ b) 7.88 y 95.99%.

25. Una fábrica de cierta marca de refresco ha tomado al azar 9 semanas del año, observando la temperatura media correspondiente en grados centígrados (X) y la cantidad de los refrescos en miles (Y) pedidos durante cada uno de dichos períodos. Los datos se resumen en la siguiente tabla:

X	28	14	12	31	30	19	24	15	16
Y	60	19	12	75	70	40	55	25	25

Hallar la recta de regresión de mínimos cuadrados para predecir la cantidad de pedidos. ¿Se puede planificar la producción en base a la temperatura?.

Rp. a) $Y = -23.9 + 3.154X$. b) $r^2 = 0.9868$.

26. Los siguientes datos son los precios de venta en dólares Y de una marca de automóviles usados X años:

X	1	2	3	4	5	6
Y	6350	5695	5790	–	4985	4890

a) Ajustar una curva de mínimos cuadrados de la forma $Y = AB^X$.
b) Estimar el precio de venta de un automóvil que tenga 4 años de uso.

Rp. a) $Y = 6559.78(0.95)^X$, $r = -0.9658$. b) $\hat{y} = 5{,}342.98$

27. Los siguientes datos son la inversión neta (X) y la tasa de interés (Y)

X	12	8	10	7	6	5	5
Y	4	5	6	7	8	9	10

Se plantean dos modelos para relacionar Y con X

$$Y = AX^B \quad \text{e} \quad Y = a + bX$$

¿Qué modelo se ajusta mejor a los datos?, ¿por qué?

Rp. Modelo no lineal, $Y = 40.29(X)^{-0.9086}$, $r = 0.93$, $r^2 = 0.87$, Modelo lineal, $Y = 12.63 - 0.74X$ $r = 0.91$, $r^2 = 0.82$. Mejor ajusta el modelo no lineal.

28. Ajustar por el método de mínimos cuadrados una curva de la forma:

$$Y = \frac{1}{A + BX}$$

a los siguientes pares de datos:

X	4	8	12	16	20	24	28	32
Y	24	21	20	15	14	10	7	5

Rp. $Y' = 1/Y$. $Y' = 0.003146 + 0.005176X$, $r = 0.9127$

29. Ajustar por el método de mínimos cuadrados una curva de la forma:

$$Y = 5 + \frac{1}{A + BX}$$

a los datos del problema 19. (hacer primero $Y' = Y - 5$).

30. La presión P (kg./cm^2.) de un gas correspondiente a diferentes volúmenes V (cm^3.) se registró en la siguiente tabla:

Ajustar a los datos una curva de mínimos cuadráticos de la forma: $PV^B = A$ y estimar P cuando V es 110 cm.^{3}.

V	50	60	70	80	90	100
P	79.7	65.1	52.6	36.8	25.7	8.7

Rp. $Y' = 6.84 + 2.83X'$, $PV^{2.83} = 6918309.709$, 11.56, $r = 0.908$

31. El número Y de bacterias por unidad de volumen presentes en un cultivo de X horas se registró en la siguiente tabla:

X	0	1	2	3	4
Y	20	35	50	80	190

Ajustar a los datos una curva de mínimos cuadráticos de la forma:

$$Y = -10 + A.B^X$$

y estimar el número de bacterias a las 5 horas.

Rp. a) $Y = -10 + 27.7(1.57)^X$, 254, $r = 0.98$.

Apéndice 4.1

4.4 NUMEROS INDICES

4.4.1 Introducción

Uno de los métodos estadísticos que se utilizan con mayor frecuencia en economía, administración de empresas, demografía y otros campos de la estadística aplicada , es el de números índices.

Básicamente un *número índice*, es el cociente de cualquier medición de una variable (o más variables) con respecto a una de sus mediciones que se toma como base.

El objetivo de los números índices es ***cuantificar variaciones*** de las mediciones de una variable a través del tiempo. En este sentido el número índice es el cociente de la medición de la variable en un período determinado con respecto a un *período base*.

Las mediciones pueden estar relacionadas con ***cantidad, precio* o *valor***.

Los números índices se clasifican en ***índices simples o elementales* e *índices compuestos o agregados***.

El número **índice simple** se calcula a partir de una sola variable. Mientras que un **índice compuesto** se calcula a partir de dos o más variables.

Los índices compuestos se clasifican en *índices no ponderados* e *índices ponderados*.

4.4.2 Indices simples

Definición. Sea x_t la medición de una variable cuantitativa X registrada en un período determinado t (año, mes o día) y x_0 $(x_0 \neq 0)$ la medición de la variable para el período base t_0. Se denomina *índice simple* de X para el período t con

respecto al período base t_0, al número que denotaremos por I_{t/t_0} o $I_{t/t_0}(X)$ o I_t y que se define por:

$$I_{t/t_0}(X) = \frac{x_t}{x_0}, \qquad \text{o,} \qquad I_{t/t_0}(X) = \frac{x_t}{x_0} \times 100\%.$$

El *porcentaje de variación* entre los valores x_0 y x_t se calcula por:

$$\%\text{ variación } = \left(\frac{x_t}{x_0} - 1.00\right)\% = (I_{t/t_0} \times 100 - 100)\%.$$

Si el porcentaje de variación es positivo se dice que ha habido un incremento, si es negativo se dice que ha habido una baja.

EJEMPLO 4.8.

En el cuadro 4.6 se dan los promedios de los salarios, en dólares, de los trabajadores de una empresa, de 1975 a 1983. Calcular los correspondientes números índices para cada uno de los nueve años utilizando como año base:

a) 1975, b) 1978, c) 1983.

Cuadro 4.6

		í n d i c e s		
Años	salarios	1975=100	1978=100	1983=100
1975	310	100.0	81.6	54.4
1976	330	106.5	86.8	57.9
1977	370	119.4	97.4	64.9
1978	380	122.6	100.0	66.7
1979	430	138.7	113.2	75.4
1980	450	145.2	118.4	78.9
1981	480	154.8	126.3	84.2
1982	540	174.2	142.1	94.7
1983	570	183.9	150.0	100.0

SOLUCION.

a) La primera columna de índices del cuadro 4.6 se obtuvo dividiendo cada cifra anual entre 310, que es el salario del año base 1975 (1975=100) lo que es lo mismo multiplicar cada cifra anual por su recíproco: 100/300=0.323 .

La **interpretación** del número índice es como sigue. El índice 183.9, por ejemplo, significa que en 1983 ha habido un aumento respecto a 1975 de $(183.9 - 100)\% = 83.9\%$. Por otra parte, si la cifra resultante es negativa se dice que ha habido una baja.

Similarmente se calcularon los índices de las otras dos columnas, tomando como base los años b) 1978 y c) 1983 respectivamente.

4.4.3 Indices simple de precios, de cantidades, y de valores

Definición. Si p_t y p_0 son los valores de la variable *precio, P*, en los períodos respectivos t y t_0, entonces, el *índice simple de precios* en el período t con respecto al período base t_0, es el número:

$$I_{t/t_0}(P)=\frac{p_t}{p_0},\ \text{o},\ I_{t/t_0}(P)=\frac{p_t}{p_0}\times 100\%.$$

Definición. Si q_t y q_0 son los valores de la variable *cantidad, Q*, en los períodos respectivos t y t_0, entonces, *el índice simple de cantidades* en el período t con respecto al período base t_0, es el número:

$$I_{t/t_0}(Q)=\frac{q_t}{q_0},\ \text{o},\ I_{t/t_0}(Q)=\frac{q_t}{q_0}\times 100\%.$$

Definición. El **valor**, *V*, de cierto artículo en un período determinado es igual *a su precio multiplicado por la cantidad vendida* (o producida). El **índice de valor simple** en un período determinado t con respecto al período base t_0, se define por:

$$I_{t/t_0}(V)=\frac{p_t q_t}{p_0 q_0}=I_{t/t_0}(P)\times I_{t/t_0}(Q).$$

Donde $p_t q_t$ y $p_0 q_0$ son los valores respectivos en el período t y en el período base.

EJEMPLO 4.9.

En la tabla 4.7 se dan los precios promedios en dólares y las cantidades de consumo promedios en kilogramos de un artículo desde 1980 a 1982. Tomando como base el año 1980, calcular los índices de precios, de cantidades, y de valores para 1981 y 1982.

SOLUCION.

Indices de precios:

$$I_{81/80}(P)=\frac{p_{81}}{p_{80}}=\frac{20}{15}=1.33\ 6133\% \qquad I_{82/80}(P)=\frac{p_{82}}{p_{80}}=\frac{25}{15}=1.67\ \text{ó}\ 167\%$$

Indices de cantidades:

$$I_{81/80}(Q)=\frac{q_{81}}{q_{80}}=\frac{7.4}{6.5}=1.14\ \text{ó}\ 114\%, \quad I_{82/80}(Q)=\frac{q_{82}}{q_{80}}=\frac{7.8}{6.5}=1.20\ \text{ó}\ 120\%$$

Indices de valores:

$$I_{81/80}(V)=\frac{p_{81}q_{81}}{p_{80}q_{80}}=\frac{20\times 7.4}{15\times 6.5}=151.8 \quad I_{82/80}(V)=\frac{p_{82}q_{82}}{p_{80}q_{80}}=\frac{25\times 7.8}{15\times 6.5}=2$$

Cuadro 4.7 Indice de precios, cantidad y valor

Año	Precio $	Cantidad Kg.	Indice Precios	Indice Cantidad	Indice Valor
1980	15	6.5	100	100	100
1981	20	7.4	133	114	152
1982	25	7.8	167	120	200

4.4.4 Indices compuestos o agregados

Definición. Un número índice compuesto se define como una combinación de números índices simples cada uno de ellos referidos a una misma base.

Los índices compuestos se clasifican en compuestos no ponderados y compuestos ponderados.

4.4.4.1 Indices compuestos no ponderados.

Uno de los métodos de cálculo de índices compuestos no ponderados o simples, es el método de la ***media agregada simple*** conocida también como índice agregado simple.

Definición. El ***índice agregado simple*** es el cociente de la suma de las medidas de dos o más variables en el período t entre la suma de las medidas de esas variables en el período base t_0.

El **índice agregado simple de precios** de varios artículos en un período t con respecto al período base t_0 se define por:

$$I_{t/t_0}(P)=\frac{\Sigma p_t}{\Sigma p_0}.$$

El **índice agregado simple de cantidades** de varios artículos en un período t con respecto período base t_0 se define por:

$$I_{t/t_0}(Q)=\frac{\Sigma q_t}{\Sigma q_0}.$$

NOTA. El índice compuesto no ponderado de valor se define por:

$$I_{t/t_0}(V) = \frac{\sum p_t q_t}{\sum p_0 q_0}.$$

EJEMPLO 4.10.

En el cuadro 4.8 se da una canasta de artículos básicos que comprende 4 ítems A, B, C, y D, los precios en dólares y las cantidades consumidas en kilogramos durante los años 1980 y 1985. Tomando el año 1980 como base, calcular los índices compuestos no ponderados de precios y cantidades de 1985.

SOLUCION.

Tomando el año 1980 como base, el índice agregado simple de cantidades para 1985 resulta:

$$I_{1985/1980}(P) = \frac{\sum p_t}{\sum p_0} = \frac{0.8 + 0.9 + 1.50 + 4.50}{0.50 + 0.80 + 1.00 + 3.00} = \frac{7.7}{5.3} = 1.45.$$

En consecuencia, de 1980 a 1985 hubo un aumento del 45% en el precio de esa canasta de artículos.

Tomando el año 1980 como base, el índice agregado simple de cantidades para 1985 resulta:

$$I_{1985/1980}(Q) = \frac{\sum q_t}{\sum q_0} = \frac{6.2 + 5.0 + 2.0 + 3.0}{5.5 + 4.0 + 1.5 + 2.4} = \frac{16.2}{13.4} = 1.21.$$

En consecuencia, de 1980 a 1985 hubo un aumento del 21% de artículos de esa canasta.

Cuadro 4.8 Cálculo compuestos no ponderados.

	Precio promedio		Cantidad consumo		Indices	
Ítem	1980 (p_0)	1985(p_t)	1980(q_0)	1985(q_t)	Precios	Cantidad
A	0.50	0.80	5.5	6.2	1.600	1.127
B	0.80	0.90	4.0	5.0	1.125	1.250
C	1.00	1.50	1.5	2.0	1.500	1.333
D	3.00	4.50	2.4	3.0	1.500	1.250
	5.30	7.70	13.4	16.2	5.725	4.960

NOTA. Otra forma de definir el índice compuesto no ponderado es el *método de la media aritmética simple:*

$$\frac{\sum \text{Indices}}{\text{nú mero de índices.}}$$

Por ejemplo, por este método, el índice de precios de 1985 con base en 1980, (cuadro 4.8), es $5.725/4 = 1.43$. Mientras que el índice de cantidades de 1985 con base en 1980 es $4.960/4 = 1.24$.

Los índices compuestos no ponderados asignan igual importancia a cada precio o cantidad componente, esto permite que un bien con un precio o valor alto, domine el índice. Por esta razón no es muy utilizado.

4.4.4.2 Indices compuestos ponderados: de Laspeyres, de Paasche y de Fisher

Las ponderaciones usadas para índices compuestos de precios son las cantidades de los bienes o ítems.

Las ponderaciones para el índices de cantidades agregadas son los precios de los bienes del año base.

Definición. El **índice de precios de Laspeyres** en un período t con respecto a un período base t_0, es la media aritmética ponderada de los índices simples de precios p_t/p_0 que usa como ponderación a los valores del año base p_0q_0, esto es,

$$IL_{t/t_0}(P) = \frac{\sum p_0 q_0 \dfrac{p_t}{p_0}}{\sum p_0 q_0} = \frac{\sum p_t q_0}{\sum p_0 q_0}$$

Definición. El **índice de cantidades de Laspeyres** en un período t con respecto a un período base t_0, es la media aritmética ponderada de los índices simples de cantidades q_t/q_0 que usa como ponderación a los valores del año base p_0q_0, esto es,

$$IL_{t/t_0}(Q) = \frac{\sum p_0 q_0 \dfrac{q_t}{q_0}}{\sum p_0 q_0} = \frac{\sum p_0 q_t}{\sum p_0 q_0}$$

Definición. El **índice de precios de Paasche** en un período t con respecto a un período base t_0, es la media aritmética ponderada de los índices simples de precios p_t/p_0 que usa como ponderación a los valores del año base p_0q_t, esto es,

$$IP_{t/t_0}(P)=\frac{\sum p_0q_t\dfrac{p_t}{p_0}}{\sum p_0q_t}=\frac{\sum p_tq_t}{\sum p_0q_t}.$$

Definición. El **índice de cantidades de Paasche** en un período t con respecto a un período base t_0, es la media aritmética ponderada de los índices simples de cantidades q_t/q_0 que usa como ponderación a los valores del año base p_tq_0, esto es,

$$IP_{t/t_0}(Q)=\frac{\sum p_tq_0\dfrac{q_t}{q_0}}{\sum p_tq_0}=\frac{\sum p_tq_t}{\sum p_tq_0}.$$

NOTA. $IL_{t/t_0}(P)\times IP_{t/t_0}(Q)=\dfrac{\sum p_tq_0}{\sum p_0q_0}\times\dfrac{\sum p_tq_t}{\sum p_tq_0}=\dfrac{\sum p_tq_t}{\sum p_0q_0}=I_{t/t_0}(V).$

Definición. El **índice ideal de Fisher,** es la media geométrica de los índices de Laspeyres y Paasche.

El *índice ideal de precios de Fisher* en el período t es:

$$IF_{t/t_0}(P)=\sqrt{IL_{t/t_0}(P)\times IP_{t/t_0}(P)}.$$

El *índice ideal de cantidades* de Fisher en el período t es:

$$IF_{t/t_0}(Q)=\sqrt{IL_{t/t_0}(Q)\times IP_{t/t_0}(Q)}.$$

EJEMPLO 4.11

La tabla 4.9 contiene las unidades, precios promedios y consumos *per capita* de 3 artículos básicos en una ciudad en los períodos de 1980 y 1985. Calcular los índices compuestos: a) de precios y b) de cantidades, por los métodos de Laspeyres, Paasche y Fisher del período 1985, tomando como base el año 1980.

Tabla 4.9 Precios y cantidades de tres ítems

Item	Unidades	Precio promedio		Cantidad consumo	
		1980(p_0)	1985(p_t)	1980(q_0)	1985(q_t)
A	litro	\$10	\$15	40	60
B	pieza	15	20	80	100
C	docena	20	25	20	40

SOLUCION.

Construyendo la tabla 4.10 para los cálculos, resultan:

a) El índice de precios de Laspeyres en el período 1985 tomando a 1980 como año base:

$$IL_{85/80}(P) = \frac{\sum p_t q_0}{\sum p_0 q_0} = \frac{600 + 1{,}600 + 500}{400 + 1{,}200 + 400} = \frac{2{,}700}{2{,}000} = 1.35 \text{ o } 135\%.$$

El índice de precios de Paasche en el período 1985 tomando a 1980 como año base:

$$IP_{85/80}(P) = \frac{\sum p_t q_t}{\sum p_0 q_t} = \frac{900 + 2{,}000 + 1{,}000}{600 + 1{,}500 + 800} = \frac{3{,}900}{2{,}900} = 1.34 \text{ o } 134\%.$$

El índice ideal de precios de Fisher de 1985, base 1980

Cuadro 4.10 Cálculo de índices compuestos ponderados

Item	Precios por cantidades			
	$p_0\ q_0$	$p_t\ q_0$	$p_0\ q_t$	$p_t\ q_t$
A	400	600	600	900
B	1,200	1,600	1,500	2,000
C	400	500	800	1,000
	2,000	2,700	2,900	3,900

$$IF_{85/80}(P) = \sqrt{IL_{85/80}(P) \times IP_{85/80}(P)} = \sqrt{1.35 \times 1.34} = 1.34 \text{ o } 134\%.$$

De estos resultados, se concluye que en el año 1985 hubo un incremento en el precio de esa canasta del 35% según el método de Laspeyres, del 34% según los métodos de Paasche y de Fisher.

b) El índice de cantidades de Laspeyres en el período 1985 tomando a 1980 como año base es

$$IL_{85/80}(Q) = \frac{\sum p_0 q_t}{\sum p_0 q_0} = \frac{600 + 1{,}500 + 800}{400 + 1{,}200 + 400} = \frac{2{,}900}{2{,}000} = 1.45 \text{ o } 145\%.$$

El índice cantidades de Paasche en el período 1985 tomando a 1980 como año base es

$$IP_{85/80}(Q) = \frac{\sum p_t q_t}{\sum p_t q_0} = \frac{900 + 2{,}000 + 1{,}000}{600 + 1{,}600 + 500} = \frac{3{,}900}{2{,}700} = 1.44 \text{ o } 144\%.$$

El índice ideal de cantidades de Fisher es

$$IF_{85/80}(Q) = \sqrt{1.45 \times 1.44} = 1.44 \quad \text{o } 144\%.$$

De estos resultados, se concluye que, en el año 1985, hubo un incremento en la cantidad de consumo de esa canasta del 45% según el método de Laspeyres, del 44% según los métodos de Paasche y de Fisher.

4.4.5 Cambio del período base

Es frecuente cambiar la base de un número índice en un período dado a un período más reciente para reemplazar los índices que se encuentran obsoletos con el fin de que las comparaciones actuales resulten más significativas.

Para cambiar el índice $I_{t/a}$ de base antigua, a, al índice $I_{t/n}$ de base nueva, n, se utiliza la siguiente regla:

$$I_{t/n} = \frac{I_{t/a}}{I_{n/a}}, \quad \text{o expresado en \%}, \; I_{t/n} = \frac{100}{I_{n/a}} \times I_{t/a}.$$

EJEMPLO 4.12.

En la tabla 4.11 se dan los índices A tomando a 1980 como año base. Obtener los índices

a) B tomando como base nueva al año 1990.

b) C tomando como base nueva al año 1985.

SOLUCION.

a) El índice B_1 de 1980 en base 1990 es:

$$I_{80/90} = \frac{100}{I_{90/80}} \times I_{80/80} = \frac{100}{450} \times 100 = 22.22.$$

El índice B_2 de 1985 en base 1990 es:

$$I_{85/90} = \frac{100}{I_{90/80}} \times I_{85/80} = \frac{100}{450} \times 300 = 66.67.$$

Tabla 4.11 Cambio de base

Año	Indice A	Indice B	Indice C
1980	100	B_1=22.22	C_1=33.33
1985	300	B_2=66.67	C_2=100
1990	450	B_3=100.0	C_3= 150

El índice B_3 de 1990 en base 1990 es:

$$I_{90/90} = \frac{100}{I_{90/80}} \times I_{90/80} = \frac{100}{450} \times 450 = 100.$$

b) El índice C_1 de 1980 en base 1985 es:

$$I_{80/85} = \frac{100}{I_{85/80}} \times I_{80/80} = \frac{100}{300} \times 100 = 33.33.$$

El índice C_2 de 1985 en base 1985 es:

$$I_{85/85} = \frac{100}{I_{85/80}} \times I_{85/80} = \frac{100}{300} \times 300 = 100.$$

El índice C_3 de 1990 en base 1985 es:

$$I_{90/85} = \frac{100}{I_{85/80}} \times I_{90/80} = \frac{100}{300} \times 450 = 150.$$

4.4.6 Empalme o fusión de dos series de números índices

Con frecuencia una serie de números índices sufre cambios por adición de ciertos productos o exclusión de otros, así como por cambio del período base obteniéndose una nueva serie de números índices. El problema es fusionar ambas series de números índices a partir de un nuevo período base.

Para fusionar dos series distintas de números índices y formar una serie nueva de números índices, esta nueva serie debe tener un índice de **empalme, fusión o traslape** para las dos series, de manera que se puedan calcular ambos tipos de índices para ese año de traslape.

Para retroceder los índices de la serie nueva, cada índice de la serie antigua se convierte en un índice de la serie nueva dividiendo el índice de empalme (100),

entre el índice antiguo de la base nueva, luego multiplicando por el índice antiguo. Esto es,

$$\text{Indice nuevo} = \frac{100}{\text{Indice antiguo de la base nueva}} \times \text{Indice antiguo.}$$

También, **para avanzar** los índices de la serie antigua cada número índice de la serie nueva se convierte en un índice de la serie antigua dividiendo el índice antiguo de la base nueva entre el índice de empalme (100), luego multiplicando por el índice nuevo. Esto es,

$$\text{Indice antiguo} = \frac{\text{Indice antiguo de la base nueva}}{100} \times \text{Indice nuevo.}$$

EJEMPLO 4.13.

En la tabla 4.12, la segunda columna muestra los índices antiguos desde 1980 hasta 1982 con respecto al año base 1980. La tercera columna muestra los nuevos índices de 1982 a 1984 con año base 1982. Efectúe el empalme de los índices con base a) 1982 y b) 1980.

SOLUCION.

En el año 1982 aparecen dos índices, el antiguo 120 y el nuevo 100, este período elegimos como empalme.

a) En la columna de índices empalmados con base a 1982, primero se obtiene el coeficiente $k = 100/120 = 0.833$. Luego, el índice para 1980 es $0.833(100) = 83.3$, el índice para 1981 es $0.833(110) = 91.63$.

b) En la columna de índices empalmados con base a 1980, el coeficiente es igual a $k = 120/100 = 1.20$. En consecuencia el índice para 1983 es $1.20(130) = 156$, y el índice para 1984 es $1.20(140) = 168$.

Tabla 4.12 Empalme de dos series de índices

Año	Indice antiguo base 1980	Indice nuevo base 1982	Indice empalmado base 1982	Indice empalmado base 1980
1980	100	--	83.3	100
1981	110	--	91.6	110
1982	120	100	100	120
1983	--	130	130	156
1984	--	140	140	168

4.4.7 Uso de los números índices

Los números índices se usan con frecuencia para cuantificar las diferencias de los valores de una o más variables a lo largo del tiempo. Las entidades estatales o privadas confeccionan números índices para diversos fenómenos económicos.

Uno de los usos más importantes de los números índices es la denominada **deflación de precios e ingresos.** Esta técnica es el proceso de "ajustar" precios e ingresos y expresarlos en términos del valor de la moneda de un período base.

Para la deflación estadística se usa como **deflactor** el índice simple de precios al consumidor *IPC, indicador del costo de vida.* Este índice simple, como ya es sabido, se obtiene dividiendo el precio p_t en un período t, entre el precio p_0 en un período base t_0. Esto es,

$$IPC = \frac{p_t}{p_0} \times 100.$$

El *valor* (precio o salario) *deflacionado* o *valor real* en el período t con respecto al período base t_0, se obtiene dividiendo el valor nominal del período t entre el *IPC* de ese período, ,esto es,

$$\text{valor deflacionado } = \frac{\text{valor nominal}}{IPC} \times 100.$$

Si el valor es el precio, se tendrá el *precio real o deflacionado* y si el valor es el ingreso o salario, se tendrá *el salario real o deflacionado.*

Por **ejemplo**, si el salario de una persona fue de 4,000 unidades monetarias en 1990 y de 8,000 unidades monetarias en 1991, entonces hubo un incremento nominal de $4,000. Pero, si el índice de precio al consumidor fue de 1.2 en 1991 con respecto a 1990, entonces su salario real es de 8,000/1.2 = 6,666.67 dólares, y ha tenido un incremento real de sólo 2,666.67 dólares.

NOTA.

Si los *IPC* van decreciendo en una serie de tiempo, los salarios nominales de un período t se *indexan*, de manera que no pierdan su poder adquisitivo con respecto a un período base a $(a < t)$, multiplicando el salario nominal del período a por IPC_t/IPC_a. Por ejemplo, si con respecto al año base 1989, el IPC de 1990 fue de 110%, el de 1991 fue de 120% y si el salario nominal en 1990 fue de 1300, entonces, el salario nominal en 1991 de manera que no pierda el poder adquisitivo de 1990 es:

$$\text{Salario nominal de } 1991 = \frac{120}{110} \times 1{,}300 = 1{,}418.18.$$

* **El índice de salarios reales**, *ISR*, para un determinado período *t*, en base al período t_0, se calcula utilizando la fórmula:

$$ISR_{t/t_0} = \frac{SR_t}{SR_0} \times 100.$$

Siendo SR_t el salario real en el período *t* y SR_0 el salario real en el período base t_0.

* **El poder adquisitivo** del dinero, *PA*, es el recíproco del índice de precios al consumidor. Esto es,

$$PA = (1/IPC) \times 100\%.$$

* **El índice del poder adquisitivo** del dinero, *IPA*, es el cociente entre el índice de precios del año base, IPC_0, que se toma como base de comparación, y el índice de precios al consumidor en el período *t*, IPC_t. Esto es,

$$IPA = \frac{IPC_0}{IPC_t} \times 100.$$

***La pérdida del poder adquisitivo**, *PPA*, o el porcentaje de *desvalorización* del dinero, se cuantifica por:

$$PPA = (1 - \frac{IPC_0}{IPC_t}) \times 100.$$

Por **ejemplo**, si el *IPC* fue de 1.25 en 1980 tomando como base a 1979, entonces, el índice del poder adquisitivo del dinero en ese año es:

$$IPA = 1.00/1.25 = 0.8, \text{ o, } 80\%.$$

Entonces, la pérdida del poder adquisitivo es la cifra, $PPA = 20\%$.

EJEMPLO 4.14.

En la tabla 4.13 se dan los salarios nominales (en dólares) y los índices de precios al consumidor de 1980 a 1986 con base a 1980, en una determinada ciudad. Para cada período y en base a 1980, calcular: a) Los salarios reales, b) los índices de salarios reales, c) los índices del poder adquisitivo.

Cuadro 4.13

Años	Salario	*IPC*	*SR* salario/*IPC*	*ISR* *SR*/75	*IPA* 100/*IPC*
1980	75.00	100	75.00	100.0	100.0
1981	80.45	110	73.14	97.52	90.9
1982	93.18	120	77.65	103.53	83.3
1983	104.25	145	71.90	95.87	69.0
1984	130.35	190	68.60	91.47	52.6
1985	170.75	250	68.30	91.07	40.0
1986	200.00	280	71.43	95.24	35.7

SOLUCION.

a) Los salarios reales , *SR*, de la cuarta columna se obtienen dividiendo cada salario nominal entre su respectivo índice de precios al consumidor, *IPC*.
b) Los índices de salarios reales , *ISR*, de la quinta columna se obtienen dividiendo cada salario real entre el salario del año base que es igual a 75 dólares.
c) Los índices de poder adquisitivo, *IPA*, de la sexta columna se obtienen dividiendo el *IPC* = 100 del año base, entre cada uno de los *IPC* de la tercera columna, y expresado luego en porcentajes.

4.4.8 Tasas anuales y mensuales: Inflación.

Uno de los problemas prácticos que pese a su sencillez, se presta a dificultades, es el de la determinación de las tasas de variación (aumento o disminución) anual o mensual de un índice.

Definición. Sean I_t, $t = 1,2,...$ los índices medidos en los períodos (meses o años): 1,2,...,t,... respectivamente. Las **tasa o índice de variación** del índice correspondiente al período *t* se define por:

$$IV_t = \left(\frac{I_t - I_{t-1}}{I_{t-1}}\right)\times 100 = \left(\frac{I_t}{I_{t-1}} - 1\right)\times 100.$$

Definición. (Inflación). Si en la fórmula anterior los índices son de precios al consumidor (*IPC*), la tasa de variación se denomina *inflación*.

Por **ejemplo**, si en diciembre de 1982 el *IPC* fue 653.1 y en diciembre de 1983 el *IPC* fue de 796.1, entonces, el porcentaje de variación de precios (inflación anual) de 1983 es:

$$\frac{796.1 - 653.1}{653.1}\times 100 = 21.896.$$

Definición. El **índice de variación acumulado** del período *a* al período *t*, se define por:

$$IVA_{t/a} = \left[\left(\frac{IV_t}{100}+1\right)\left(\frac{IV_{t-1}}{100}+1\right)...\left(\frac{IV_a}{100}+1\right)\right]\times 100 - 100.$$

siendo de mayor aplicación cuando el período es mensual.

EJEMPLO 4.15

En la tabla 4.14 se dan los índices de precios al consumidor de diciembre de 1979 a setiembre de 1980, calcular, a) la inflación mensual, b) la inflación mensual acumulada de diciembre de 1979 a setiembre de 1980.

SOLUCION.

a) Las inflaciones mensuales se dan en la tercera columna. La inflación, por ejemplo, de Marzo de 1980, es igual a,

$$\frac{120-114}{114}\times 100\% = 5.26\%.$$

b) Las inflaciones acumuladas se dan en la cuarta columna. La inflación acumulada, por ejemplo, de Enero (a) a marzo (t) de 1980 es,

$$IVA_{t/a} = \left[\left(\frac{5.26}{100}+1\right)\left(\frac{7.55}{100}+1\right)\left(\frac{6.00}{100}+1\right)\right]\times 100 - 100 = 19.9995\% = 20\%.$$

Cuadro 4.14 Computo de la inflación mensual y acumulada

Meses	Indice Dic. 1979=100	Inflación mensual *IV*	Inflación mensual *IVA* Acumulada
Dic. 1979	100	---	----
Ene 1980	106	6.00	6.00
Febrero	114	7.55	14.00
Marzo	120	5.26	20.00
Abril	130	8.33	29.99
Mayo	150	15.38	49.98
Junio	168	12.00	67.97
Julio	180	7.14	79.96
Agosto	195	8.33	94.95
Setiembre	218	11.79	117.93

4.4.9 Devaluación

La **devaluación** es la pérdida del valor o del poder adquisitivo externo del dinero que se cuantifica generalmente con el precio del dólar.

Sea C_t el nuevo tipo de cambio y C_a el antiguo tipo de cambio, el porcentaje de **aumento del tipo de cambio** se define por:

$$\%\ \text{aumento} = \left(\frac{C_t}{C_a}-1\right)\times 100\%$$

Por ejemplo, si el tipo de cambio de 2.08 soles por dólar sustituye a uno de 2.02 soles por dólar, entonces el porcentaje de alza del dólar es

$$\%\ \text{alza} = \left(\frac{2.08}{2.02}-1\right)\times 100\% = 2.97\%$$

Es decir, hay un aumento en el precio del dólar de 2.97%

El porcentaje de **devaluación** se define por

$$\% \text{ devalución } = \left(1 - \frac{C_a}{C_t}\right) \times 100\%$$

Por ejemplo, si el tipo de cambio de 2.08 soles por dólar sustituye a uno de 2.02 soles por dólar, entonces el porcentaje devaluación del sol es

$$\% \text{ devaluación} = \left(1 - \frac{2.02}{2.08}\right) \times 100\% = 2.88\%$$

Es decir, el aumento de 2.97% en el precio del dólar produce una devaluación en el sol de 2.88%.

EJERCICIOS

1. Dados los promedios anuales de los precios al por menor (por kilogramos) de una artículo de consumo diario de 1980 a 1985

Años	1980	1981	1982	1983	1984	1985
Precios	4.8	5.4	5.7	6.6	6.3	5.2

Calcular los índices de precios tomando como base el año:
a) 1980, b) 1982.

Rp. a) en %, 100, 112.5, 118.75, 137.75, 131.25, 108.33 b) en %, 84.21, 94.74, 100, 115.79, 110.53, 91.23

2. Dadas las siguientes producciones de un bien en kg.

Año	1985	1986	1987	1988	1989	1990	1991	1992
Producción	40	36	41	94	90	104	107	106

a) Calcular los índices de producción usando como base el año 1985.
b) Cambiar la base a 1989.

3. Si en 1991 el precio de un bien de consumo fue 12% más que en 1990, 25% más que en 1992 y 10% menos que en 1989, calcule los índices de precios de estos cuatro años con base en 1989.

Rp. Los precios e índices (base 1989) de 1989 a 1992 son respectivamente: $1.24x$, x, $1.12x$, $0.896x$, y 100, 8.65, 90.3, 72.25.

4. Si en 1992 la producción de un bien fue de 15% menos que el año anterior, 30% más que en 1993 y 40% menos que en 1990, calcule los índices de cantidades de consumo de dicho bien de 1990 a 1993 con base en 1992.

Rp. Las cantidades e índices (base 1992) de 1990 a 1993 son respectivamente: $1.42x$, x, $0.85x$, $0.6538x$, y 167, 117.65, 100 y 76.92.

5. Los índices de producción de un bien de 1989 a 1993 fueron respectivamente 1.00, 1.15, 1.20, 1.24 y 1.30. Si en 1992 se produjeron 2.79 toneladas de dicho bien, calcular la producción de los demás años.

Rp. 2.25, 2.5875, 2.7, 2.79 y 2.925

6. En enero de 1990 una fábrica gastó en salarios de 50 obreros, la suma de $ 6,000 y en diciembre del mismo año con 75 obreros gastó $9,900. Tomando como base el mes de enero, calcular

a) el índice de empleados para diciembre (cantidad).
b) el índice de gastos de salarios (valor).
c) el índice de costo por obrero o índice *per capita* (de precio).

Rp. a) 75/50 = 1.50, b) 9,900/6000 = 1.65, c) $IpIq = Iv$, $Ip = 1.65/1.50 = 1.10$

7. Si el índice de cantidad del año 1990 con base de 1980 es de 110 y con base de 1985 es de 130, calcular el índice de cantidad de 1985 con base de 1980.

Rp. 0.8462 o 84.62%

8. En 1994 el precio de venta de un bien se ha fijado en 20% más con respecto al año anterior. Si se quiere que el valor total de la venta se incremente en 80% durante el año 1994 con respecto al año anterior, ¿en que porcentaje debe incrementarse la cantidad de bienes a vender?.

Rp. $I(p)I(q) = I(v)$, $I(q) = 1.80/1.20 = 1.5$. debe incrementarse en 50%

9. En 1993 el valor de la producción de un bien disminuyó en 20% con respecto al de 1992, mientras que el precio aumentó en 30%, ¿cuánto es el porcentaje de variación de la producción?.

Rp. $I(q) = 0.6154$, disminuyó 38.46%.

10. Una fábrica bajó la producción de un bien en 15% de 1993 con respecto a 1992, pero aumentó el precio del mismo en 25% en el mismo lapso, calcular el porcentaje de variación del valor total de las ventas de dicho bien, en 1993 con respecto a 1992. Rp. aumentó en 6.25%

11. Si en el año 1990 el ingreso de una persona ha tenido un incremento del 110% con respecto al año 1985 que era de \$10,000 y si por otra parte el índice del costo de vida se triplicó en ese lapso, a) ¿cuánto es el ingreso real de esa persona en 1990?, b) cuánto debería ser su ingreso nominal de manera que mantenga su poder adquisitivo de 1985?.

Rp. a) 21,000/3 = 7,000, es el 70% de lo que percibía en 1985, b) \$30,000

12. Los índices de precios *ligados* o en *cadena* (Esto es, índices que tienen base variable igual al período inmediatamente anterior) de 1991 con 1990 es de 1.20 y de 1992 con 1991 es de 1.40 y de 1993 con 1992 es 1.50. Calcular el índice de precios de 1993 con base de 1990.

Rp. $I_{93/90}=I_{91/90}\times I_{92/91}\times I_{93/92}=1.2\times 1.4\times 1.5=2.52$

13. En la tabla que sigue se dan los precios en soles al por mayor y la producción en kilogramos de tres artículos básicos *A*, *B* y *C* de los años de 1985 a 1987 en una determinada región.

a) Calcular el índice de precios i) media , ii) mediana para 1987 con base en 1985.
b) Calcular el índice compuesto de precios de 1987 con base en 1985, por el método de i) Laspeyres, ii) Paasche, y iii) Fisher.
c) Calcular el índice compuesto de cantidad de 1987 con base en 1985, por el método de i) Laspeyres, ii) Paasche, y iii) Fisher.

	precios			Cantidad		
Artículos	1985	1986	1987	1985	1986	1987
A	2.4	2.8	3.6	3400	4000	4300
B	32	44	55	60	66	75
C	12	18	25	30	33	37

Rp. ai) El índice promedio aritmético de precios para 1987 en base a 1985 es igual a 1/3(3.6/2.4 + 55.0/32.0 + 25/12) = 1.77 o 177%. aii) El índice mediana de precios es 1.72 o 172%, b) 1.5603, 1.5596, 1.5599, c) 1.2609, 1.26028, 1.2605.

14. En 1990 el precio de un artículo, expresado en dólares de 1985, fue de \$0.60. Si el índice de precios creció en un 20% en ese lapso, ¿cuál es el precio nominal de ese artículo en 1990?. Rp. 0.72\$

15. El índice de precios al consumidor respecto a 1989 fue de 120% para 1990 y de 140% para 1991. ¿Qué salario nominal debe tener una persona en 1991, si en 1990 fue de \$400?.

Rp. El salario nominal en el período t conociendo el salario nominal del período m es:
$SN_t = (SN_m/IPC_m)\times IPC_t$, $\quad SN_{91} = (SN_{90}/IPC_{90})\times IPC_{91}$,
$SN_{91} = (400/120)\times 140 = 466.67$.

16. Los ingresos medios por persona (ingresos per cápita) y los índices de precios al consumidor de 1977 a 1984 se dan en la tabla que sigue.

a) Calcular los aumentos reales y nominales con base 1977.
b) ¿Cuáles deben ser los aumentos nominales de modo que el poder adquisitivo se mantenga igual al del año 1977?.

Año	1977	1978	1979	1980	1981	1982	1983	1984
Ingresos	700	800	850	900	1000	1200	1300	1500
I.P.C (%)	100	105	110	120	125	130	140	145

Rp. a) aumentos nominales: 0, 100. 150, 200, 300, 500, 600 y 800 aumentos reales: 0, 61.9. 72.73, 50, 100, 223, 228.57 y 334.48. b) 0. 35. 70, 140. 175, 210. 280 y 315.

17. Los salarios anuales de miles de dólares de un empleado y los índices de precios al consumidor I.P.C. de 1980 a 1985 se dan en la siguiente tabla:

Año	1980	1981	1982	1983	1984	1985
Salarios	30	35	37	40	42	50
I.P.C(%)	100	120	130	140	160	167

a) Hallar los salarios reales de manera que queden expresados en dólares de 1980 (deflacionar los salarios).
b) Replantear el salario de 1980 en dólares de cada año (indexar salarios).
c) Determine el poder adquisitivo de compra del dólar para cada uno de tales años en términos del dólar del año base.

Rp. a) 30, 29.2, 28.46, 28.57, 26.25, 29.94, b) 30, 36, 39, 42, 48, 50.1 c) 1, 0.83, 0.77, 0.71, 0.625, 0.599.

18. Dadas las series de índices de precios al consumidor A y B:

Año	1983	1984	1985	1986	1987	1988	1989	1990
Indice A, (%)	100	110	130	135	–	–	–	–
Indice B, (%)	–	–	–	100	115	120	135	140

Empalmar los dos series de índices utilizando como año base
a) 1986, b) 1983.

Rp. a) 74.1, 81.48, 96.3, 100, 115, 120, 135, 140, b) 100, 110, 130, 135, 155.25, 162, 182.25. 189.

19. Dadas las series de índices de precios al consumidor A y B de 1983 a 1990 de la tabla que sigue:

Año	1983	1984	1985	1986	1987	1988	1989	1990
Indice A, (%)	200	300	350	450	–	–	–	–
Indice B, (%)	–	–	–	90	100	130	150	200

a) Empalmar los dos conjuntos de índices utilizando como año base el año 1986.
b) Calcular la inflación para cada año.

c) Si en 1989 el sueldo promedio de un sector de la industria fue de $20,000 y el de 1985 fue de $12,000, ¿en cuál de los períodos el sueldo promedio fue mejor?.

Rp. a) Indices, en %, B, base 1986: 100, 111.1, 144.4, 166.67, 222.2. Indices empalmados: 44.4, 66.67, 77.78, 100, 111.1, 144.4, 166.67, 222.2. c) en 1985.

20. Se dispone de las siguientes series: I.P.1, I.P.2 de índices para el precio de venta de un bien:

Año	1988	1989	1990	1991	1992	1993
I.P.1.	1	1.25	2.5	5	–	–
I.P.2.	–	–	–	0.75	1	1.5

a) Si el precio de venta de tal bien fue de $30 en 1993, ¿Cuánto hubiera sido su precio de venta en 1990?.

b) Determine los índices de precios de venta de 1988 a 1993 con base a 1992.

Rp. a) 7.5, b) 0.15, 0.1625, 0.375, 0.75, 1, 1.5.

21. Si la tasa de inflación del mes dos con respecto al mes uno fue de 2.4% y la del mes tres con respecto al mes dos fue de 2%. Calcule la inflación del mes uno al mes tres. Rp. 4.448.

22. Los índices en porcentajes de precios al consumidor de 1985 a 1990 fueron respectivamente: 80, 86.7, 100, 112, 120 y 128.

a) Calcule la tasa de inflación de cada año.

b) Si para cierto sector laboral el salario promedio era de $200 en 1986 y de $300 en 1989, ¿en cuál de los períodos el salario promedio fue mejor?.

Rp. a) 8.375, 15.34, 12, 7.14, y 6.67 b) En 1986, 200/0.867 =$230 y en 1989, 300/1.2 = $250.

Capítulo 5

PROBABILIDAD

5.1 Experimento aleatorio, espacio muestral, eventos

5.1.1 Experimento aleatorio

En estadística la palabra *experimento* se utiliza para describir un proceso que genera un conjunto de datos cualitativos o cuantitativos. En la mayoría de los casos, los resultados del experimento dependen del azar, por lo tanto no pueden pronosticarse con exactitud.

Definición. Un experimento aleatorio es todo proceso que consiste de la ejecución de un *acto* (o *prueba*) una o más veces, cuyo resultado en cada prueba depende del azar y en consecuencia no se puede predecir con certeza.

Por ejemplo, son experimentos aleatorios: lanzar un dado y observar el resultado, contar objetos defectuosos producidos diariamente por cierto proceso, aplicar una encuesta para obtener opiniones, etc..

5.1.2 Espacio muestral

Definición. Se denomina **espacio muestral** al conjunto que consiste de todos los resultados posibles de un experimento aleatorio. Este conjunto se denotará por Ω.

Cada resultado posible de un experimento aleatorio es un elemento del espacio muestral. A cada elemento del espacio muestral se denomina también *punto muestral*. Esto es, el espacio muestral se describe por

$$\Omega = \{\omega / \omega \text{ es un punto muestral}\}.$$

Si el espacio muestral tiene un número finito de elementos es posible enlistar a todos estos, y si el número de elementos es grande o infinito el espacio muestral se describirá mediante un enunciado o regla.

EJEMPLO 5.1.
A continuación se dan algunos experimentos aleatorios y sus correspondientes espacios muestrales:

1) El experimento aleatorio de lanzar un dado y observar el resultado obtenido, es de una sola prueba, cuyo espacio muestral se puede escribir como el siguiente conjunto de puntos muestrales:

$$\Omega_1 = \{1, 2, 3, 4, 5, 6\}.$$

2) El experimento aleatorio de lanzar una moneda 3 veces, consiste de 3 pruebas, cuyo espacio muestral puede escribirse como el conjunto de ternas ordenadas:

$$\Omega_2 = \{CCC, CCS, CSC, SCC, SSC, SCS, CSS, SSS\}.$$

NOTA. Los espacios muestrales de experimentos aleatorios que consisten de dos o más pruebas sucesivas se obtienen también de un **diagrama tipo árbol**, como el de la figura 5.1, para Ω_2.

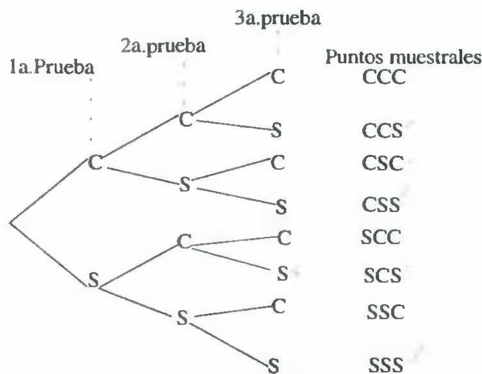


Fig. 5.1 Diagrama del árbol.

3) Si el experimento aleatorio es lanzar una moneda y un dado a la vez, y observar los resultados, el espacio muestral es el conjunto:

$$\Omega_3 = \{1C, 2C, 3C, 4C, 5C, 6C, 1S, 2S, 3S, 4S, 5S, 6S\}.$$

4) Si el experimento aleatorio es lanzar una moneda tantas veces como sea necesario hasta que aparezca la primera cara, su espacio muestral es el conjunto:

$$\Omega_4 = \{C, SC, SSC, SSSC, \ldots etc.\}.$$

5) Si el experimento aleatorio es medir la vida útil (en horas) de una marca de artefacto eléctrico, su espacio muestral es el conjunto:

$$\Omega_5 = \{t \in \Re \,/\, t \geq 0\}.$$

(Aquí, $\Re$ representa a los números reales).

6) Si el experimento aleatorio consiste en determinar la posición de caída de un dardo que es tirado hacia un blanco circular de radio 5 cm., su espacio muestral es el conjunto:

$$\Omega_6 = \{(x, y) \in \Re^2 \,/\, x^2 + y^2 \leq 25\}.$$

Clasificación de los espacios muestrales. Por el número de elementos o puntos muestrales, los espacios muestrales se clasifican en:

1) **Discretos finitos,** consisten de un número finito de elementos, por ejemplo, los espacios Ω_1, Ω_2, Ω_3.

2) **Discretos infinitos,** consisten de un número infinito numerable de elementos, por ejemplo, el espacio Ω_4.

3) **Continuos,** consisten de un número infinito no numerable de elementos, por ejemplo, los espacios Ω_5, y Ω_6

5.1.3 Eventos

Definición. Se denomina **evento** a cualquier subconjunto de un espacio muestral.

Por ejemplo, si $\Omega = \{w_1, w_2, ..., w_n\}$ es un espacio muestral finito de n elementos, en él se pueden definir 2^n eventos diferentes, entre ellos:

* El **evento imposible,** $\varnothing$, que no tiene puntos muestrales, en consecuencia *no ocurre* nunca .
* Los **eventos unitarios o elementales,** $\{w_i\}$, que contienen un sólo punto muestral.
* Los **eventos compuestos,** que consisten de dos o más eventos.
* El **evento seguro o cierto,** Ω, el mismo espacio muestral, ya que es el subconjunto que contiene a todos los eventos elementales.

Definiciones:

1. Se dice que **un evento *A* ocurre,** si contiene por lo menos un punto muestral de algún experimento aleatorio. Esto es,

   Un evento A **ocurre** *si y sólo si* existe $w \in A$.

2. Un **evento *A* no ocurre** *si y sólo si* $w \notin A$.

3. El evento A es un **subevento** o está **contenido** en el evento B, simbolizado, $A \subset B$, si toda vez que ocurre A ocurre también B.

4. Los eventos A y B son **iguales,** $A = B$, si y sólo si $A \subset B$ y $B \subset A$.

5. Se denomina **complemento** del evento A al evento que se denota por A^C o A' o $\overline{A}$, que consiste de todos los puntos muestrales que no están en el evento A (figura. 5.2.a), esto es,

$$A^c = \{w \in \Omega / w \notin A\}.$$

El evento A^c describe el evento de que no ocurra A.

Operaciones con eventos

1. Se denomina **unión** de los eventos A y B, al evento $A \cup B$ que consiste de todos los puntos muestrales que pertenecen a A o a B, o a ambos (figura. 5.2 b), esto es,

$$A \cup B = \{w \in \Omega / w \in A \vee w \in B\}.$$

El evento $A \cup B$ describe el evento de que **ocurra por lo menos uno de ellos**.

2. Se denomina **intersección** de los eventos A y B al evento $A \cap B$ que consiste de todos los puntos muestrales que son comunes a A y a B (figura 5.2 c), esto es,

$$A \cap B = \{w \in \Omega / w \in A \wedge w \in B\}.$$

El evento $A \cap B$ describe el evento de que **ocurran ambos A y B**.

3. Dos eventos A y B son **mutuamente excluyentes** o **disjuntos,** si no tienen elementos en común (figura 5.2d), esto es, si $A \cap B = \varnothing$.

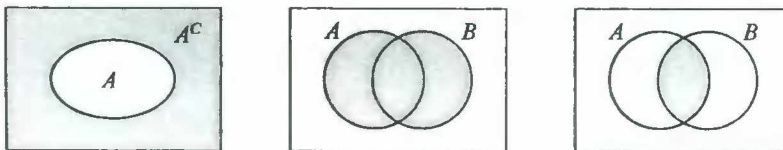


Fig. 5.2. a) Complemento **Fig.5.2. b)** Unión **Fig.5.2. c)** Intersección

En general, diremos que los eventos: $A_1, A_2, ..., A_n$ son mutuamente excluyentes si,

$$A_i \cap A_j = \varnothing, \quad \forall i \neq j, \; i, j = 1,2,3,...,n$$

4. La **diferencia** del evento A menos B es el evento $A-B$, que consiste de todos los puntos muestrales que pertenecen al evento A y no pertenecen al evento B (figura. 5.2e), esto es,

$$A-B=\{w\in\Omega / w\in A\wedge w\notin B\}.$$

El evento $A-B=A\cap B^C$ **describe el evento de que ocurra A y no ocurra B**.

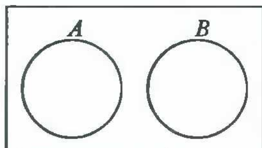


Fig. 5.2 d) Disjuntos

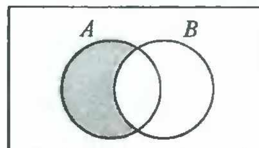


Fig. 5.2 e) Diferencia

5. El producto cartesiano de los eventos A y B, es el evento A×B que consiste de todos los pares ordenados de puntos muestrales (w_1, w_2), siendo $w_1\in A$ y $w_2\in B$, esto es,

$$A\times B=\{(w_1,w_2)/ w_1\in A\wedge w_2\in B\}.$$

El evento $A\times B$ **describe el evento de que ocurra 1° A y 2° B**.

El espacio muestral Ω asociado a un experimento aleatorio compuesto de dos experimentos aleatorios Ω_1 y Ω_2, cuyos espacios muestrales respectivos son Ω_1 y Ω_2, se puede expresar como un producto cartesiano, esto es,

UNIVERSO → $\Omega=\Omega_1\times\Omega_2$

Por **ejemplo**, si se lanzan 2 monedas, a la vez o en forma consecutiva, y si Ω_1 y Ω_2 son los espacios muestrales de las monedas 1 y 2 respectivamente, entonces, el espacio muestral Ω del experimento compuesto es:

$$\Omega=\Omega_1\times\Omega_2=\{C,S\}\times\{C,S\}=\{(C,C),(C,S),(S,C),(S,S)\}.$$

Algebra de eventos

Las siguientes identidades básicas se verifican para eventos

1) $A\cup A=A$, $A\cap A=A$
2) $A\cup B=B\cup A$, $A\cap B=B\cap A$
3) $A\cup A^C=\Omega$, $A\cap A^C=\varnothing$
4) $A\cup\varnothing=A$, $A\cap\varnothing=\varnothing$

5) $A \cup \Omega = \Omega$, $\quad A \cap \Omega = A$

6) $\Omega^C = \varnothing$, $\quad \varnothing^C = \Omega$, $\quad (A^C)^C = A$

7) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

8) $(A \cup B)^C = A^C \cap B^C$, $\quad (A \cap B)^C = A^C \cup B^C$

NOTA.
Sea $A_1, A_2, ..., A_n$ cualquier colección finita de n eventos,

1. El evento de que ocurra por lo menos uno de ellos se describe por el conjunto:

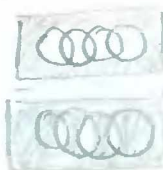
$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_i .$$

2. El evento de que ocurran todos ellos juntos se describe por el conjunto:

$$A_1 \cap A_2 \cap ... \cap A_n = \bigcap_{i=1}^{n} A_i .$$

3. La regla de De Morgan,

$$\left[\bigcup_{i=1}^{n} A_i\right]^c = \bigcap_{i=1}^{n} (A_i)^c, \text{ y } \quad \left[\bigcap_{i=1}^{n} A_i\right]^c = \bigcup_{i=1}^{n} (A_i)^c .$$



4. El producto cartesiano de los n eventos es el evento:

$$A_1 \times A_2 \times ... \times A_n = \{(w_1, w_2, ..., w_n) / w_i \in A_i \ \forall i = 1,2,...,n\}.$$

Este evento se utiliza para describir los resultados de un experimento aleatorio compuesto de n experimentos aleatorios simples, por ejemplo, tirar un dado n veces o tirar n dados a la vez.

5.2 Conteo de puntos muestrales

Cuando es grande el número de resultados posibles de un experimento aleatorio, no suele ser fácil el recuento de tales resultados, por eso, es necesario dar ciertas reglas que nos faciliten el conteo de puntos muestrales.

5.2.1 Número de puntos muestrales

El número de elementos de un evento arbitrario A se denota por $n(A)$ o por $\#(A)$. Es evidente que:

$$n(\varnothing)=0 \quad \text{y} \quad n(A)\geq 0 \text{ , para todo evento } A.$$

Regla de la multiplicación.

" Si una operación puede realizarse de n_1 formas y por cada una de estas una segunda operación puede realizarse de n_2 formas, entonces, las dos operaciones pueden realizarse de $n_1 \times n_2$ formas".

Esto es, si A y B son dos conjuntos finitos, entonces, el número de elementos del producto cartesiano $A \times B$ está dado por

$$n(A)\times n(B)$$

Por **ejemplo,** ¿cuántos elementos tiene el espacio muestral que resulta de lanzar dos dados a la vez?.

Si A es el evento "resultados del dado 1", entonces, $n(A) = 6$ y si B es el evento: "resultados del dado 2", entonces, $n(B)=6$. Luego, el número de resultados posibles al lanzar los dos dados es igual a:

$$n(A)\times n(B)=6\times 6=36.$$

La extensión de la regla de la multiplicación con k experimentos se puede expresar en términos de conjuntos como sigue:

Si $A_1, A_2, \cdots, A_k$ son k conjuntos finitos, entonces, el número de elementos del producto cartesiano $A_1 \times A_2 \times \cdots \times A_k$ es igual a:

$$n(A_1)\times n(A_2)\times \cdots \times n(A_k).$$

Por **ejemplo**, ¿cuántos números pares de tres dígitos distintos se pueden formar con los dígitos: 1, 4, 7, 8, 9?.

Si A_1, A_2, y A_3, representan los dígitos de centenas, decenas y unidades respectivamente, entonces, $n(A_3)=2$, $n(A_2)=4$, $n(A_1)=3$. Luego, el total de números es: $n(A_1 \times A_2 \times A_3) = 3\times 4\times 2 = 24$

Por dar otro **ejemplo**, suponga que de una urna que contiene n objetos numeradas de 1 a n se escogen k de ellos una a una. Sea A_i el evento que denota los posibles resultados de la extracción i. $i=1,2,\ldots,k$. El número de formas de extracción es igual al número de elementos del evento $A_1 \times A_2 \times \ldots \times A_k$, esto es,

$$n(A_1)\times n(A_2)\times \ldots \times n(A_k)=\underbrace{n\times n\times \ldots \times n}_{k \text{ veces}}=n^k,$$

si la extracción es con reposición, y es igual a:

$$n(A_1) \times n(A_2) \times ... \times n(A_k) = n \times (n-1) \times (n-2) \times ... \times (n-k+1),$$

si la extracción es sin reposición.

Cada elemento $(a_1, a_2, ..., a_k)$ del evento $A_1 \times A_2 \times ... \times A_k$ se denomina en el primer caso, **variación con repetición,** y en el segundo caso, **variación sin repetición**.

Regla de la adición.

1. Si A y B son eventos mutuamente excluyentes, entonces,

$$n(A \cup B) = n(A) + n(B).$$

Por ejemplo, el experimento de lanzar una moneda o un dado tiene $2 + 6 = 8$ resultados posibles.

2. En general, si los n eventos $A_1, A_2, ..., A_n$, no ocurren mutuamente juntos (son disjuntos mutuamente), entonces,

$$n(A_1 \cup A_2 \cup ... A_n) = \sum_{i=1}^{n} n(A_i).$$

3. Si A y B son dos eventos cualesquiera, entonces,

$$n(A \cup B) = n(A) + n(B) - n(A \cap B).$$

4. Utilizando 3) se verifica que si A, B, y C son tres eventos cualesquiera,

$$n(A \cup B \cup C) = n(A) + n(B) + n(C) - n(AB) - n(AC) - n(BC) + n(ABC).$$

5.2.2 Variaciones

A) Variaciones simples

Definición. Se denominan variaciones simples o sin repetición o simplemente variaciones de k objetos tomados de n objetos distintos, a cada uno de los arreglos u órdenes que se hagan con los k objetos, de manera, que estos arreglos difieran en algún elemento o en el orden de colocación.

Por ejemplo, las variaciones de 2 elementos del conjunto $A = \{a, b, c, d\}$ son los siguientes arreglos:

ab, ac, ad, ba, bc, bd, ca, cb, cd, da, db, dc.

El **número de variaciones** diferentes de k objetos tomados de n objetos distintos, denotado por V_k^n o ${}_nV_k$, está dado por:

$$V_k^n = n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!}.$$

En efecto, hay n posibilidades para la primera posición, $(n-1)$ para la segunda, ...etc.,.. y, $(n-k+1)$ posibilidades para la k-ésima posición. Luego se aplica el principio de la multiplicación.

Por **ejemplo**, el número de variaciones de 2 objetos tomados de 4 objetos diferentes es:

$$V_2^4 = 4 \times 3 = \frac{4!}{(4-2)!} = 12.$$

NOTA: Una fórmula de recurrencia para el total de variaciones (para programar variaciones) es la expresión siguiente:

$$V_k^n = (n-k+1)V_{k-1}^n.$$

B) Variaciones con repetición

Definición. Se denominan **variaciones con repetición** de k objetos tomados de n objetos distintos, a cada uno de los arreglos de k de tales objetos de manera que dos, tres, ..., k de ellos, puede ser uno mismo de los n objetos.

Por ejemplo, las variaciones con repetición de dos elementos tomados de los elementos a, b, c son:

aa, ab, ac, ba, bb. bc, ca, cb. cc.

El número de variaciones con repetición de k objetos a partir de n objetos distintos, que denotaremos por VR_k^n, es:

$$VR_k^n = \underbrace{(n)(n)...(n)}_{k \text{ veces}} = n^k$$

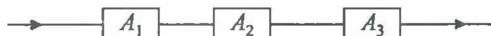
Por ejemplo, el número de variaciones con repetición de dos elementos tomados de 3 elementos distintos es:

$$VR_2^3 = 3 \times 3 = 3^2 = 9.$$

4. Un lote de N artículos contiene k defectuosos, describir el espacio muestral del número de artículos extraídos hasta obtener el último defectuoso.

Rp. $\Omega=\{k, k+1, k+2, \ldots, N\}$.

5. a) Describa el espacio muestral del funcionamiento del siguiente sistema:



donde $A_i=1$ si está operativo, $A_i=0$ si está descompuesto, $i=1, 2, 3$.

b) Describa los eventos A:" por lo menos una componente funciona", y B: "todo el sistema funciona". ¿Son A y B mutuamente excluyentes?.

Rp. $\Omega=\{(x_1,x_2,x_3)/x_i=0,1,\ i=1,2,3,\ \}$, b) no.

6. Un experimento consiste en observar la vida útil de dos objetos, describa el evento "la duración del primero más la duración del segundo es al menos 4 años". Rp. $\Omega=\{(x, y)\ /\ x+y\geq 4\}$.

7. En un edificio de 10 pisos entran al ascensor al primer piso 3 personas. Cada una baja al azar a partir del segundo piso. ¿De cuántas maneras estas personas pueden bajar en pisos diferentes?.

Rp. $9\times8\times7=504$.

8. Cierto producto se arma en tres etapas. En la primera hay 5 líneas de ensamblado, en la segunda 6, y en la tercera 4. ¿De cuántas maneras distintas puede hacerse circular el producto durante el proceso de ensamblado?.

Rp. $5\times6\times4$.

9. Una caja contiene 8 dulces de piña, 6 de naranja y 4 de fresa. ¿Cuántos elementos tiene el espacio muestral que resulta de extraer al azar un dulce de cada sabor?.

Rp. $C_1^8C_1^6C_1^4=192$.

10. Cinco alumnos forman cola en la ventanilla de la secretaría de la facultad.

a) ¿De cuántas maneras diferentes pueden hacer la cola?.

b) ¿De cuántas maneras si el más alto debe estar al comienzo?.

c) ¿De cuántas maneras si el más alto y el más bajo deben estar en extremos opuestos?.

d) ¿De cuántas maneras si el más alto y el más bajo no deben estar juntos?.

Rp. a) $5!$, b) $4!$, c)$3!2$, d) $5!-4!2!$

11. Cierta marca de sierra eléctrica es calificada por especialistas, en cuanto a rendimiento, como: "Muy buena", (B1); o, "buena", (B2); o "regular", (B3), y

en cuanto al precio, como "cara", (C1), o "barata"; (C2). ¿De cuántas maneras es calificada la sierra eléctrica por los especialistas?.

Rp. 6.

12. Un vendedor de automóviles acaba de recibir un embarque de 15, de los cuales 10 son del modelo A y 5 del modelo B. ¿De cuántas maneras puede vender 4 de los automóviles,

a) si los 4 son del mismo modelo?.

b) si dos son del modelo A?.

c) si al menos uno es del modelo B?.

Rp. a) $C_4^{10} + C_4^5$, b) $C_2^{10} C_2^5$, c) $\sum_{k=1}^4 C_k^5 C_{4-k}^{10}$.

13. Suponga que una urna contiene 10 fichas de color blanco, 10 de color rojo, 10 de color amarillo y 10 de color negro. Las fichas del mismo color van numeradas de 1 a 10. Un experimento consiste en extraer al azar una de las fichas de la urna. Dados los eventos: A: "color blanco", B: "número menor que 4", y C: "número par", determinar el número de puntos muestrales de los siguientes eventos compuestos:

$$A \cap B,\ A \cap C,\ A \cap B \cap C,\ A \cap B^c,\ A^c \cap B,\ A^c \cap B^c \cap C,\ A^c \cap B^c \cap C^c$$

$$A^c \cap B \cap C^c,\ \ A \cap B^c \cap C^c, A^c \cup B^c \cup C^c.$$

14. ¿De cuántas maneras diferentes pueden colocarse en un estante 6 libros de Matemática, 2 de Historia y 4 de Lógica si los libros de la misma materia deben estar juntos y si,

a) no se distingue entre los libros de la misma materia?

b) se distingue entre los libros de la misma materia?

Rp. a) $P_{6,2,4}^{12} = \frac{12!}{6! \times 2! \times 4!} = 13.860$, b) $(6! \times 2! \times 4!)3! = 207.360$

15. De 8 hombres y 7 mujeres ¿cuántos comités de 10 miembros se pueden formar si cada uno de ellos debe contener cuando menos 5 mujeres?.

Rp. $C_5^8 C_5^7 + C_4^8 C_6^7 + C_3^8 C_7^7 = 1772$

16. ¿Cuántas parejas se pueden elegir de 4 hombres y 6 mujeres si cierto varón rehusa tener como pareja a dos de las mujeres?.

Rp. $C_1^4 + C_1^3 C_1^6$.

17. Seis hombres y seis mujeres compiten realizando cierta tarea. Si los seis primeros puestos son ocupados por 4 hombres y dos mujeres determine el número de casos. Rp. $6!C_4^6 C_2^6 = 162{,}000$

18. ¿Cuántas sumas diferentes de dinero se pueden formar con cuatro monedas cada una de distinto valor?. Rp. $\sum_{i=1}^{4} C_i^4 = 2^4 - 1$

19. Una compañía desea ascender a 3 de sus 10 gerentes a posiciones de Vicepresidente de ventas, de manufacturas y de finanzas. Hallar el número de formas distintas de efectuar los ascensos. Rp. V_3^{10}

20. ¿De cuántas maneras diferentes puede un padre dividir 8 regalos entre sus 3 hijos, si el mayor debe recibir 4 regalos y los menores 2 cada uno?.

Rp. $C_4^8 C_2^4 C_2^2 = 420.$

21. Dados los conjuntos A de 4 elementos y B de 8 elementos, ¿cuántos conjuntos de 6 elementos se pueden formar si cada conjunto debe contener:
a) sólo un elemento de A?, b) cuando menos un elemento de A?.

Rp a) 224, b) 896.

22. Hallar el número de formas diferentes en que se pueden hacer:
i) una selección , ii) un ordenamiento con 4 letras de las palabras:
a) *Cloroformo*,
b) *Eliminación*.

23. Un estudiante debe contestar 5 de 7 preguntas de un examen, ¿de cuántas maneras diferentes puede escoger las cinco
a) sin ninguna restricción?.
b) si las dos primeras son obligatorias?.
c) si debe contestar 3 de las 4 primeras?.

Rp. a) C_5^7, b) C_3^5, c) $C_3^4 C_2^3$

24. Un estudiante planea matricularse en los cursos A, B, y C. Los horarios de A son a las 8, 11 y 15 horas. Los de B son a las 8, 10 y 15 horas y los de C a las 10, 12 y 15 horas. Si las clases son de una hora, ¿cuántos horarios distintos puede preparar en los 3 cursos de manera que no haya cruces?.

Rp. ABC: 2, ACB: 3, BAC: 4, BCA: 4, CAB: 1, total 14

25. ¿De cuántas formas pueden instalarse en línea 5 focos blancos y 6 focos rojos si deben colocarse a) alternativamente?, b) los blancos juntos?.

Rp. a) $5!\times6!$, b) $7!\times5!$.

26. Un microbús tiene 29 asientos para pasajeros, distribuidos en 6 filas de 4 asientos cada uno, con un pasillo en el medio y al final 5 asientos juntos. ¿De cuántas maneras diferentes podrán ubicarse 25 pasajeros de modo tal, que los 14 asientos que dan a las ventanillas queden ocupados?.

Rp. $V_{14}^{25}V_{11}^{15}$

27. De un conjunto de n objetos numerados de 1 al n ($n \geq 4$), se seleccionan k objetos a la vez. ¿En cuantos casos,

a) todos son menores que m $(1 \leq m \leq n)$?.

b) sólo 3 son mayores a m $(k \geq 3)$?. Rp. a) C_k^{m-1}, b) $C_3^{n-m}C_{k-3}^{m}$.

28. Se tienen 40 fichas, donde hay 4 de color rojo. Si se reparten estas fichas entre 4 niños, de manera que a cada uno toque 10 fichas. ¿En cuántos casos a cada uno toca una ficha roja?. Rp. $4!\times36!/(9!)^4$.

29. Un grupo de 5 hombres y 10 mujeres, se divide al azar en 5 grupos de 3 personas cada una. Calcular el número de maneras en que cada grupo contenga un hombre.

Rp. $(C_1^5C_2^{10})(C_1^4C_2^8)(C_1^3C_2^6)(C_1^2C_2^4)(C_1^1C_2^2)$.

30. Se tiene n objetos iguales numerados de 1 a n, $n>3$.

a) Si se permutan los n objetos, ¿de cuántas maneras k $(k < n)$ de ellos ocuparán sus posiciones correspondientes?.

b) Si se tienen 3 cajas en las cuales se distribuyen los objetos, ¿de cuántas maneras diferentes se pueden distribuir a fin de que sólo una caja quede vacía?.

Rp. a) $((n-k)!)\,C_k^n$, b) $(2^n-2)C_2^3$

31. Se distribuyen k bolas numeradas de 1 a k en n $(k < n)$ cajas alineadas, ¿de cuántas maneras k cajas *vecinas* contienen una sola bola cada una?.

Rp. $(n-k+1)k!$.

5.3 Probabilidad de un evento

Definición. (Axiomática)

Sea Ω el espacio muestral asociado a un experimento aleatorio. La probabilidad de cualquier evento A de Ω, es el número real $P(A)$ que satisface los siguientes axiomas:

P1) $P(A) \geq 0$, para todo evento A.

P2) $P(\Omega) = 1$.

P3) Si A y B son dos eventos mutuamente excluyentes, entonces,

$$P(A \cup B) = P(A) + P(B).$$

De los axiomas de probabilidad resultan los teoremas que siguen:

TEOREMA 5.1.
Si $\varnothing$ es el evento imposible, entonces $P(\varnothing) = 0$.

Demostración.

Los eventos Ω y $\varnothing$ son disjuntos. Además, $\Omega = \Omega \cup \varnothing$.

Por el axioma P3, $P(\Omega) = P(\Omega) + P(\varnothing)$, de donde resulta $P(\varnothing) = 0$.

NOTA. Si $P(A) = 0$, entonces, no necesariamente $A = \varnothing$ (una justificación se halla en la sección 5.4.3, otra en el capítulo 6, sección 3).

TEOREMA 5.2.
Si A^c es el evento complementario del evento A, entonces,

$$P(A) = 1 - P(A^c) \quad \text{o} \quad P(A^c) = 1 - P(A).$$

Demostración.

Los eventos A y A^c son disjuntos. Además $\Omega = A \cup A^c$.

Por el axioma P3), $P(\Omega) = P(A) + P(A^c)$ y por P2), $1 = P(A) + P(A^c)$.

De donde resulta que $P(A^c) = 1 - P(A)$. También $P(A) = 1 - P(A^c)$.

TEOREMA 5.3
Si A y B son dos eventos tales que $A \subset B$, entonces, $P(A) \leq P(B)$

Demostración.

$A \subset B$, implica $B = A \cup (B - A)$, donde A y $B - A$ son eventos disjuntos. Por el axioma P3) se tiene, $P(B) = P(A) + P(B - A)$ o $P(B) - P(A) = P(B - A)$. Por otra parte, por el axioma P1), $P(B - A) \geq 0$. De donde resulta $P(B) - P(A) \geq 0$, o $P(A) \leq P(B)$.

NOTA. Para todo evento A, se verifica, $\varnothing \subset A \subset \Omega$, entonces,

$$P(\varnothing) \leq P(A) \leq P(\Omega),$$

consecuentemente $\qquad 0 \leq P(A) \leq 1$.

NOTA. La **probabilidad *P* es una *función* definida en eventos que son partes de** Ω y que asocia a cada evento A el número real $P(A) \in [0,1]$ y que satisface los axiomas: P1), P2) y P3).

TEOREMA 5.4 (*Regla de la adición de eventos*).
Si A y B son dos eventos cualesquiera, entonces:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

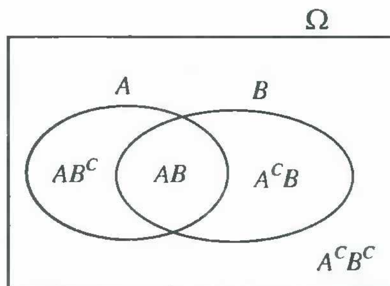


Figura 5.3

Demostración.

(Usaremos la notación AB por $A \cap B$). De la figura 5.3 se obtiene: $A \cup B = A \cup (A^c B)$, siendo, A y $A^c B$ disjuntos.

Además, $B = AB \cup (A^c B)$, siendo AB y $A^c B$ también disjuntos.
Por el axioma P3),

$$P(A \cup B) = P(A) + P(A^c B) \quad \text{y} \quad P(B) = P(AB) + P(A^c B).$$

De estas dos identidades resulta:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

NOTA. Una consecuencia inmediata del teorema 5.4 es:

$$P(A \cup B) \leq P(A) + P(B)$$

NOTA. El lector debería verificar que para tres eventos cualesquiera A, B y C, aplicando el teorema 5.4 se tiene:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

Esta regla se generaliza para n eventos (ejercicio para el lector).

EJEMPLO 5.9

Suponga que en un sorteo la probabilidad de ganar el primer premio es 2/5 y la de ganar el segundo premio es 3/8. Si la probabilidad de ganar al menos uno de los 2 premios es 3/4, calcular la probabilidad de ganar:
a) sólo uno de los dos premios, b) ninguno de los dos premios.

SOLUCION.

Sean los eventos: A: "ganar el primer premio" y B: "ganar el segundo premio". Se tiene $P(A) = 2/5$, $P(B) = 3/8$, y $P(A \cup B) = 3/4$.

Sustituyendo $P(A) = 2/5$, $P(B) = 3/8$, y $P(A \cup B) = 3/4$ en

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

resulta $$P(A \cap B) = \frac{2}{5} + \frac{3}{8} - \frac{3}{4} = \frac{1}{40}.$$

Las probabilidades de cada una de las partes de Ω se indica en la figura 5.4.

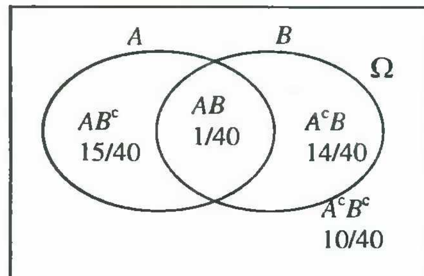


Figura 5.4

a) La probabilidad de ganar sólo uno de los dos premios es:

$$P[AB^c \text{ o } A^c B] = \frac{15}{40} + \frac{14}{40} = \frac{29}{40}$$

b) La probabilidad de no ganar ninguno de los dos premios es:

$$P(A^c B^c) = \frac{10}{40}$$

NOTA. (Extensión del axioma P3)

1. Utilizando el axioma P3), por inducción: se verifica que:
Si $A_1, A_2, \cdots, A_n$ son n eventos mutuamente excluyentes dos a dos, entonces,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \cdot$$

2. En teoría de probabilidad avanzada, en lugar del axioma P3 o su generalización, la nota anterior, se utiliza el axioma siguiente:

P4) Si $A_1, A_2, \cdots,$ es una **sucesión infinita** de eventos que se excluyen mutuamente dos a dos, entonces,

$$P(A_1 \cup A_2 \cup \ldots) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \cdot$$

5.4 Cálculo de probabilidades

5.4.1 Espacio muestral finito

Sea $\Omega = \{w_1, w_2, \cdots, w_n\}$, un espacio muestral finito de n elementos. Este espacio se puede expresar como una unión (disjunta) finita de sus eventos elementales $\{w_i\}$, esto es,

$$\Omega = \bigcup_{i=1}^{n} \{w_i\}.$$

Entonces, utilizando los axiomas P2) y P3) resulta,

$$P\left(\bigcup_{i=1}^{n} \{w_i\}\right) = \sum_{i=1}^{n} P(\{w_i\}) = 1.$$

Si todos los eventos elementales $\{w_i\}$ son **igualmente posibles** (o equiprobables) y si $P(\{w_i\}) = p$, entonces, $np = 1$, de donde, $p = \frac{1}{n}$. Consecuentemente, si A es cualquier evento del **espacio equiprobable** Ω que consiste de k $(0 \leq k \leq n)$ puntos muestrales, entonces, la probabilidad de A es el número:

$$P(A) = \underbrace{\frac{1}{n} + \frac{1}{n} + \ldots + \frac{1}{n}}_{\text{k veces}} = \frac{k}{n} = \frac{n(A)}{n(\Omega)}$$

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{\text{Casos favorables a } A}{\text{Casos posibles}}.$$

Se deja como **ejercicio** para el lector, verificar que el número $P(A) = \frac{\#(A)}{\#(\Omega)}$ **es una probabilidad**, es decir, satisface los axiomas: P1, P2 y P3.

EJEMPLO 5.10

Se lanza un dado y se observa el número obtenido. Calcular la probabilidad de obtener a) 3 puntos, b) al menos 5 puntos

SOLUCION

El espacio muestral del experimento aleatorio que consiste en lanzar un dado y observar el resultado, es el conjunto: $\Omega=\{1,2,3,4,5,6\}$

a) Si A es el evento: "obtener 3 puntos", entonces, $A=\{3\}$, y la probabilidad de A es el número:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{1}{6} = 0.167.$$

b) Si B es el evento: "obtener al menos 5 puntos", entonces, $B=\{5,6\}$, y la probabilidad de B es el número:

$$P(B) = \frac{n(A)}{n(\Omega)} = \frac{2}{6}$$

EJEMPLO 5.11.

Se lanza un dado 2 veces consecutivas. Calcular la probabilidad de obtener:

a) 7 puntos.
b) 6 puntos sólo en la segunda tirada.
c) 7 puntos ó 6 puntos sólo en la segunda tirada.
d) 7 puntos y 6 puntos sólo en la segunda tirada.

SOLUCION.

El espacio muestral del experimento aleatorio que consiste en lanzar un dado dos veces y observar los puntos obtenidos, es el conjunto:

$$\Omega = \{(i, j) / i, j = 1,2,3,4,5,6\}$$

$\#(\Omega) = 36$, este espacio tiene 36 eventos elementales equiprobables.

a) Si A es el evento, obtener "suma 7 puntos", entonces,

$$A = \{(1,6), (6,1), (2,5), (5,2), (3,4), (4,3)\}\ , \quad \#(A) = 6$$

y su probabilidad es el número:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{6}{36} = \frac{1}{6} = 0.167.$$

b) Si B es el evento "sale 6 sólo en la segunda tirada", entonces,

$$B = \{(1,6), (2,6), (3,6), (4,6), (5,6)\}, \quad \#(B) = 5$$

y su probabilidad es el número

$$P(B) = \frac{n(B)}{n(\Omega)} = \frac{5}{36} = 0.139.$$

c) El evento "sale suma 7 o sólo 6 en la segunda tirada" es:

$$A \cup B = \{(1,6), (6,1), (2,5), (5,2), (3,4), (4,3), (2,6), (3,6), (4,6), (5,6)\}$$

y su probabilidad es el número:

$$P(A \cup B) = \frac{n(A \cup B)}{n(\Omega)} = \frac{10}{36} = 0.278.$$

d) El evento "sale suma 7 y sólo 6 en la segunda tirada" es: $A \cap B = \{(1,6)\}$
y la probabilidad de este evento es el número:

$$P(A \cap B) = \frac{n(A \cap B)}{n(\Omega)} = \frac{1}{36} = 0.0278$$

EJEMPLO 5.12. (Espacio muestral no equiprobable).

Un dado se carga de tal manera que un número par tiene el doble de posibilidad de salir que un número impar. Se lanza este dado y se observa el número obtenido. Calcular la probabilidad de obtener al menos 3 puntos.

SOLUCION.

El espacio muestral del experimento aleatorio de lanzar este dado y observar el resultado obtenido es el conjunto:

$$\Omega = \{1,2,3,4,5,6\}$$

Si p es la probabilidad de obtener un número impar entonces $2p$ es la probabilidad de obtener un número par. Del axioma P2) se tiene $P(\Omega) = 1$, luego, $9p = 1$, de donde resulta $p = 1/9$.

Si A es el evento: "obtener al menos 3 puntos", entonces, $A = \{3, 4, 5, 6\}$ y la probabilidad de A es el número:

$$P(A) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} + \frac{2}{9} = \frac{2}{3} = 0.667 \cdot$$

5.4.2 Espacio muestral infinito numerable

Sea $\Omega = \{w_1, w_2, \cdots, w_n, \cdots\}$, un espacio muestral infinito numerable. Entonces, se puede expresar Ω, como una unión disjunta infinita de sus eventos elementales (equiprobables o no), esto es,

$$\Omega = \bigcup_{i=1}^{\infty} \{w_i\}.$$

Utilizando los axiomas P2 y P4) resulta:

$$P\left(\bigcup_{i=1}^{\infty} \{w_i\}\right) = \sum_{i=1}^{\infty} P(\{w_i\}) = 1.$$

Luego, si A es un evento de Ω, cuyos eventos elementales pueden ser equiprobables o no, entonces,

$$P(A) = \sum_{w_i \in A} P(\{w_i\}).$$

EJEMPLO 5.13.

Un dado se lanza sucesivamente hasta que aparezca el primer *uno*.

a) Describa el espacio muestral del experimento y determine la probabilidad de cada elemento.

b) Verificar que $P(\Omega) = 1$

c) Si dos personas A y B juegan lanzando el dado uno después del otro y si gana el que obtiene el primer ***uno***, calcular la probabilidad de que gane A el juego, si él comienza primero.

SOLUCION.

a) Sea E (éxito), el evento "sale *uno* en la tirada i", y sea F (fracaso) el evento: "no sale uno en la tirada i", donde $i = 1,2,3$, etc.

El espacio muestral que resulta de lanzar el dado sucesivamente hasta obtener un *uno* se puede escribir como el conjunto:

$$\Omega = \{E, FE, FFE, \cdots, etc\}$$

Si $w_1 = E$, $w_2 = FE$, $w_3 = FFE$, etc, entonces,

$$P(\{w_1\}) = \frac{1}{6},\ P(\{w_2\}) = \frac{5}{6} \times \frac{1}{6},\ P(\{w_3\}) = \left(\frac{5}{6}\right)^2 \frac{1}{6},$$

Resumiendo, para cada $i = 1,2,3,...$ resulta:

$$P(\{w_i\}) = \left(\frac{5}{6}\right)^{i-1} \times \frac{1}{6}$$

b) Para verificar que $P(\Omega) = 1$, se usa la suma geométrica:

$$1 + r + r^2 + \cdots = \frac{1}{1-r},\quad \text{si } |r| < 1$$

$$P(\Omega) = \sum_{i=1}^{\infty} P(\{w_i\}) = \frac{1}{6}\sum_{i=1}^{\infty}\left(\frac{5}{6}\right)^{i-1} = \frac{1}{6} \times \frac{1}{1-\frac{5}{6}} = 1 .$$

c) Si A comienza el juego, entonces, A juega las veces impares 1,3,5,...etc., y la probabilidad de que gane A es:

$$P(\text{gane A}) = \frac{1}{6} + \frac{1}{6}\left(\frac{5}{6}\right)^2 + \frac{1}{6}\left(\frac{5}{6}\right)^4 + \frac{1}{6}\left(\frac{5}{6}\right)^6 + \ldots$$

$$P(\text{gane } A) = \frac{1}{6}\left(1 + \left[\left(\frac{5}{6}\right)^2\right] + \left[\left(\frac{5}{6}\right)^2\right]^2 + \left[\left(\frac{5}{6}\right)^2\right]^3 + \ldots\right) = \frac{1}{6} \times \frac{1}{1-\left(\frac{5}{6}\right)^2} = \frac{6}{11}.$$

5.4.3 Espacio muestral continuo.

Sea A cualquier evento de un espacio muestral continuo Ω, tal que la medida (longitud o área) de A, $m(A)$, exista. Entonces, la probabilidad de A es el número,

$$P(A) = \frac{m(A)}{m(\Omega)}.$$

EJEMPLO 5.14.

La demanda de dos productos A y B varía aleatoriamente en un rango de 1000 a 5000 kilogramos. El distribuidor decide bajar el precio de venta de ambos productos si la suma de sus demandas varía de 3000 a 5000 Kg. Calcular la probabilidad de que el precio de venta de ambos productos baje.

SOLUCION.

Sea X: demanda del producto A, Y: demanda del producto B, en miles de kilogramos. El espacio muestral Ω es el conjunto:

$$\Omega = \{(x, y) \in \Re^2 / 1 \leq x \leq 5 \ \text{ y } \ 1 \leq y \leq 5\}$$

Si A es el evento " el precio de ambos productos baja", entonces, en términos de la demanda,

$$A = \{(x, y) \in \Omega / 3 \leq x + y \leq 5\}$$

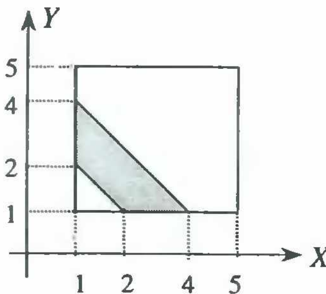


Figura 5.5

En la figura 5.5 se observa que A es la parte sombreada cuya área es:

$$Area(A) = \frac{3 \times 3}{2} - \frac{1 \times 1}{2} = 4$$

Por tanto,

$$P(A)=\frac{area(A)}{area(\Omega)}=\frac{4}{4\times 4}=0.25\cdot$$

NOTA. En el caso continuo la probabilidad de un punto en Ω es cero. Por lo tanto, $P(A)=0$, **no** implica $A=\varnothing$. En cambio, si Ω es finito, $P(A)=0$, implica $A=\varnothing$

EJERCICIOS RESUELTOS

EJERCICIO 1.

Una urna contiene 5 fichas similares de los cuales 3 son de color rojo y 2 de color azul. Si de esa urna se extraen al azar 3 fichas a la vez, calcular la probabilidad de que sólo una de ellas sea de color rojo:

a) Listando los elementos,
b) Sin listar los elementos.

SOLUCION.

a) Identifiquemos a las fichas rojas por R_1, R_2, y R_3, y a las fichas azules por A_1 y A_2.

El espacio muestral Ω es el conjunto de todas las ternas de fichas que se pueden extraer, es decir,

$$\Omega=\{ R_1R_2R_3,\ A_1R_1R_2,\ A_1R_1R_3,\ A_1R_2R_3,\ A_2R_1R_2,\ A_2R_1R_3,\ A_2R_2R_3,\ A_1A_2R_1,\ A_1A_2R_2,\ A_1A_2R_3 \}.$$

Sea A el evento "sólo una de las 3 fichas es de color rojo", entonces.

$$A=\{ A_1A_2R_1,\ A_1A_2R_2,\ A_1A_2R_3 \}.$$

La probabilidad de A es el número:

$$P(A)=\frac{n(A)}{n(\Omega)}=\frac{3}{10}.$$

b) En la mayoría de las aplicaciones no interesa saber cuales son, si no cuántos son los elementos del espacio muestral Ω y del evento A.

En este caso, el número de elementos del espacio Ω es igual al número de formas diferentes de extraer 3 fichas a la vez de la urna que contiene 5 fichas, es decir,

$$CP=C_3^5=\frac{5!}{3!2!}=10.$$

Por otra parte, el número de elementos del evento A, es igual al número de formas de extraer a la vez 3 fichas de las cuales una es roja y dos son blancas. Es decir,

$$CF = C_2^2 C_1^3 = 1 \times 3 = 3.$$

Luego, la probabilidad de A es el número $P(A)=CF/CP=3/10$.

EJERCICIO 2.

Dos alumnos se distribuyen al azar en 3 computadoras numeradas con 1, 2 y 3 respectivamente. Si ambos pueden estar en una misma computadora, pero ninguno en dos computadoras a la vez.

a) Determine los elementos del espacio muestral.

b) Calcular la probabilidad de que la computadora 3 no se ocupe.

SOLUCION.

Sean los eventos:

A_i : "alumno A en la computadora i, i = 1,2,3"

B_j : "alumno B en la computadora j, j = 1,2,3"

a) El espacio muestral Ω es el conjunto

$$\Omega = \{A_1B_1, A_1B_2, A_1B_3, A_2B_1, A_2B_2, A_2B_3, A_3B_1, A_3B_2, A_3B_3\}$$

b) Si C es el evento "la computadora 3 no se ocupa", entonces,

$$C = \{A_1B_1, A_1B_2, A_2B_1, A_2B_2\}$$

La probabilidad de C es el numero:

$$P(C) = \frac{\#(C)}{\#(\Omega)} = \frac{4}{9}.$$

Observar que $\#(\Omega) = 3 \times 3 = 9$ ya que cada alumno puede escoger cualquiera de las 3 computadoras. También $\#(C) = 2 \times 2 = 4$, ya que si la número 3 debe estar libre, entonces, cada uno puede escoger cualquiera de las dos que quedan.

NOTA.

Si la condición es que los alumnos escojan computadoras diferentes, entonces:

$$\Omega = \{A_1B_2, A_1B_3, A_2B_1, A_2B_3, A_3B_1, A_3B_2\}$$

También el evento C: "la computadora 3 no se ocupa", es

$$C = \{A_1B_2, A_2B_1\}$$

$$P(C) = \frac{\#(C)}{\#(\Omega)} = \frac{2}{6}.$$

EJERCICIO 3.

Suponga que una rifa consiste de 1000 boletos. En esta rifa un boleto se premia con \$500, dos con \$250, cinco con \$100, cien con \$5, y los demás no se premian. Si se adquiere un boleto de la rifa, calcular la probabilidad de:

a) ganar alguno de los premios,
b) ganar a lo más \$100,
c) no ganar premio alguno.

SOLUCION.

Si suponemos que los boletos están enumerados de 1 a 1000, entonces, el espacio muestral Ω que consiste en elegir un boleto al azar es el conjunto:

$$\Omega = \{1, 2, \cdots, 1000\}$$

a) Si A es el evento "ganar alguno de los premios", entonces este evento consiste de todos los boletos que serán premiados. Es decir, $\#(A) = 1 + 2 + 5 + 100 = 108$. Luego, la probabilidad de A es el número:

$$P(A) = 108/1000 = 0.108.$$

b) Si B es el evento "ganar a lo más \$100", entonces este evento consiste de todos los boletos que se premiarán con \$5 y con \$100. Es decir, $\#(B) = 5 + 100 = 105$. Luego, la probabilidad de B es el número:

$$P(B) = 105/1000 = 0.105.$$

c) Sea C el evento "no ganar premio alguno", entonces este evento consiste de todos los boletos que no serán premiados. Es decir, $n(C) = 1000 - 108 = 892$. Luego, la probabilidad de C es el número:

$$P(C) = 892/1000 = 0.892.$$

EJERCICIO 4.

Una urna contiene 13 fichas, de las cuales, 6 fichas están numeradas con 15, cuatro numeradas con 10 y tres numeradas con 5. Si de esta urna se escogen 3 fichas al azar y a la vez, calcular la probabilidad de que

a) Al menos dos de ellas tengan el mismo número
b) La suma de las tres fichas sea 30

SOLUCION.

El número de elementos del espacio muestral que consiste escoger 3 fichas a la vez de 13 es: $\#(\Omega) = C_3^{13} = 286$

a) Si A es el evento;"Al menos 2 de las 3 fichas tienen el mismo número", entonces, A^c es el evento: "las tres fichas son diferentes". El número de elementos de A^c es $\#(A^c) = C_1^6 C_1^4 C_1^3 = 72$. Luego, la probabilidad de A es el número

$$P(A) = 1 - P(A^c) = 1 - \frac{72}{286} = 0.748$$

b) Si B es el evento: "la suma de las tres fichas es 30", entonces, B consiste de una ficha numerada con 15, una numerada con 10 y una numerada con 5 o tres fichas numeradas con 10. El número de elementos de B es

$$\#(B) = C_1^6 C_1^4 C_1^3 + C_3^4 = 76\text{, y}$$

$$P(\text{B}) = \frac{\#(B)}{\#(\Omega)} = \frac{76}{286} = 0.2657$$

EJERCICIO 5.

Una lote consiste de 15 objetos, 7 de los cuales se califican como E (éxito) y el resto se califican como F (fracaso). Si se escogen 5 objetos al azar, calcular la probabilidad de que 3 sean E y 2 sean F, si se escogen:

a) Uno por uno sin reposición
b) Uno por uno con reposición
c) Los cinco a la vez

SOLUCION.

a) $CP = 15 \times 14 \times 13 \times 12 \times 11 = 360360$, ($CP = V_5^{15}$)

$$CF = C_3^5 (7 \times 6 \times 5) \times (8 \times 7) = 117600\text{, } (CF = C_3^5 V_3^7 V_2^8)$$

$$P(3E \text{ y } 2F) = \frac{CF}{CP} = \frac{117600}{360360} = 0.326$$

b) $CP = 15 \times 15 \times 15 \times 15 \times 15 = 15^5$, $CF = C_3^5 (7 \times 7 \times 7) \times (8 \times 8)$,

$$P(3E \text{ y } 2F) = \frac{CF}{CP} = \frac{C_3^5 7^3 \times 8^2}{15^5} = C_3^5 p^3 q^2,$$

donde $p = 7/15$ y $q = 8/15$

c) $CP = C_5^{15}$, $CF = C_3^7 \times C_2^8$, $P(3E \text{ y } 2F) = \dfrac{CF}{CP} = \dfrac{C_2^7 \times C_2^8}{C_5^{15}}$

NOTA. El lector debería verificar que a) y c) dan el mismo resultado.

Esto es, $$\frac{C_3^5 V_3^7 V_2^8}{V_5^{15}} = \frac{C_3^7 C_2^8}{C_5^{15}}$$

EJERCICIOS

1. Verificar que para los eventos A y B cualesquiera:

a) $P(A-B) = P(A) - P(AB)$.

b) $P(AB^c \cup A^c B) = P(A) + P(B) - 2P(AB)$.

2. Si A, B y C son eventos cualesquiera, demostrar que:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC) .$$

3. Si la probabilidad de que ocurra un evento A es 1/2 y que ocurra un evento B es 3/4, determine los posibles valores de $p = P(A \cap B)$.

Rp. 0, si $A \cap B = \varnothing$, 1/2, si $A \subset B$ y $1/4 \le p \le 1$ si $A \not\subset B$.

4. Un sistema está formado por dos componentes A y B cuyas probabilidades de falla son 1/6 y 2/15 respectivamente. Si la probabilidad de que al menos una de las dos componentes falle es 7/30, calcular la probabilidad de que:

a) ninguno de las dos componentes fallen.

b) sólo una de las componentes falle. Rp. a) 23/30, b) 1/6.

5. Un lote contiene n objetos. La probabilidad de que al menos uno sea defectuoso es 0.06, mientras que la probabilidad de que al menos dos sean defectuosos es 0.04. Calcular la probabilidad de que:

a) todos los objetos sean no defectuosos.

b) exactamente un objeto sea defectuoso. Rp. a) 0.94, b) 0.02.

6. Un monedero contiene monedas de medio sol en número igual a 4 veces el número de monedas de 20 céntimos, y contiene monedas de un sol en número igual a 3 veces el número de monedas de medio sol. Si se elige una moneda al azar, calcular la probabilidad de que su valor sea al menos de medio sol.

Rp. 16/17.

7. Como resultado de la demanda de pasajes, las líneas aéreas nacionales se han visto obligadas a aumentar el número de vuelos. Una compañía determinada tiene por el momento 5 vuelos Lima-Iquitos dos de ellos en la mañana y los otros en la tarde.

a) ¿Cuál es la probabilidad de que no haya ningún vuelo en la mañana?

b) Si se cancelan al azar dos de estos vuelos, ¿cuál es la probabilidad de que sigan habiendo un vuelo en la mañana y dos en la tarde?

Rp. a) 3/5, b) $C_1^2 C_2^3 / C_3^5$

8. Una caja contiene 5 fichas de \$10 cada una, 3 de \$30 cada una y 2 de \$50 cada una. Si se escogen 3 fichas al azar y a la vez, calcular la probabilidad de que la suma de los valores sea de \$70.

Rp. $(C_2^5 C_1^2 + C_2^3 C_1^5)/C_3^{10}$

9. Un programador de computadoras debe escoger tres trabajos de entre cinco que esperan la atención del programador. Aunque el programador no lo sabe, los trabajos varían en cuanto al tiempo de programación que requieren. Defina un espacio muestral para este experimento enumerando sus elementos. Dar una asignación de probabilidades adecuada y calcular la probabilidad de que el programador escoja los dos trabajos que requieran el menor tiempo.

Rp. Ω={123, 124, 125, 134, 135, 145, 234, 235, 245, 345}, Prob=3/10

10. Doscientas personas están distribuidas de acuerdo a su sexo y lugar de procedencia de la siguiente manera: 130 son hombres, 110 son de la capital y 30 son mujeres y de provincias. Si se eligen dos personas al azar calcular la probabilidad de que:

a) Ambos sean hombres y de provincias.

b) Al menos uno de los dos escogidos sea mujer.

Rp. a) C_2^{60}/C_2^{200}, b) $(C_1^{70}\, C_1^{130} + C_2^{70})/C_2^{200}$

11. Un lote contiene 8 artículos buenos y 4 defectuosos, si se extraen al azar 3 artículos a la vez, calcular la probabilidad de obtener por lo menos un defectuoso.

Rp. $=1-C_0^4C_3^8 / C_3^{12}$

12. Una urna contiene 20 fichas similares de las cuales 10 son rojas, 6 son azules, y 4 son verdes. Si se extraen 10 fichas al azar y a la vez, calcular la probabilidad de que

a) cinco fichas sean rojas.

b) cinco sean rojas, 3 azules; y 2 sean verdes.

Rp. a) $C_5^{10}C_5^{10}/C_{10}^{20}$, b) $C_5^{10}\,C_3^{6}C_2^{4}/C_{10}^{20}$.

13. Un comerciante tiene 12 unidades de cierto artículo de los cuales 4 tienen algún tipo de defecto. Un cliente pide para comprar 3 de tales artículos pero que no tengan defectos. Si el comerciante escoge al azar y de una sola vez 4 de tales artículos, ¿cuál es la probabilidad de que con las 4 unidades escogidas satisfaga el pedido del cliente?.

Rp. $(C_3^8C_1^4 + C_4^8)/C_4^{12}$.

14. De 6 alumnos de ingeniería y 4 de ciencias se deben seleccionar dos de ellos para hacer cierta tarea, ¿cuál es la probabilidad de que la selección esté formada por uno de ciencias y otro de ingeniería si un determinado alumno de ciencias no puede hacer pareja con 2 de ingeniería?.

Rp. $(C_1^4 + C_1^3C_1^6)/C_2^{10}$.

15. Un jurado de 7 jueces van a decidir la inocencia o culpabilidad de un reo. Suponga que 4 votan por inocencia y los otros tres por culpabilidad. Si se seleccionan al azar 3 jueces y se pregunta por su voto , ¿cuál es la probabilidad de que la mayoría de los jueces de la muestra estén a favor de la inocencia del reo?.

Rp. $(C_2^4C_1^3 + C_3^4C_0^3)/C_3^7$

16. Una mujer se quejó de discriminación de parte de una compañía local, por el hecho de ser mujer. El jurado que vio el caso estaba formado por 5 mujeres y 3 hombres y votó 5 a 3 a favor de la demandante. El abogado de la compañía apeló alegando parcialidad de los miembros de jurado de acuerdo a su sexo. ¿cuál es la probabilidad de que realmente esto haya sido cierto?

Rp. 5M3H, 5/(5+3). Sea 5M cinco mujeres votan a favor $P(5M) = C_5^5C_0^3/C_5^8$

17. Una caja contiene 16 pernos de los cuales 8 no tienen defectos, 5 tienen defectos leves, y 3 tienen defectos graves. Si se eligen 3 pernos al azar y de una sola vez. calcular la probabilidad de que los tres no tengan defectos leves.

Rp. $1-(C_3^5/C_3^{16})$.

18. Suponga que 3 alumnos se matriculan al azar en un curso que tiene cinco secciones: H1, H2, H3, H4, y H5, pudiendo los tres matricularse en una misma sección. Calcular la probabilidad de que ninguno de ellos se matricule en la sección H1. Rp. $4^3/5^3$.

19. Una urna contiene doce fichas de las cuales tres están premiadas. Si a una persona le toca extraer cinco fichas al azar y a otra persona el resto, ¿cuál es la probabilidad de que las tres fichas premiadas sean obtenidas por una misma persona?.

Rp. $(C_2^9/C_5^{12})+(C_4^9/C_7^{12})=(C_2^9+C_4^9)/C_5^{12}$.

20. Se diseña un circuito que debe tener 8 resistencias numeradas de 1 a 8 conectadas en serie. Si se instalan cuatro resistencias de marca A y cuatro de marca B, ¿cuál es la probabilidad de que las cuatro resistencias de marca A

a) ocupen las cuatro primeras posiciones?.

b) estén siempre juntas? Rp. a) $4!\times 4!/8!$, b) $5!\times 4!/8!$

21. Cien personas fueron encuestadas acerca de sus preferencias sobre tres productos A, B, y C. Se encontró que 50 prefieren el A, 37 el B, y 30 el C. Además 12 prefieren A y B, 8 sólo A y C, 5 sólo B y C, y 15 sólo C. De cinco personas encuestadas elegidas al azar, calcular la probabilidad de que 2 de ellas prefieran B y C, 2 sólo A y B, y una prefiera los tres productos

Rp. $C_2^7 C_2^{10} C_1^2/C_5^{100}$.

22. En una producción de 10,000 artículos , 1,000 de estos pueden tener al menos uno de 3 tipos de defectos A, B y C de la siguiente manera 650 de A, 372 de B, 590 de C, 166 de A y B, 434 de A y C, 126 de B y C. Si un artículo de esta producción es elegido al azar, calcular la probabilidad de que tenga,

a) los 3 tipos de defectos, b) sólo un tipo de defecto.

Rp. a) 0.0114, b) 0.0502.

23. Un dado normal se tira cinco veces, calcular la probabilidad de obtener:

a) todos los resultados diferentes, b) al menos dos resultados iguales.

Rp. a) $V_5^6/6^5$, b) $1-(V_5^6/6^5)$

24. Calcular la probabilidad de que de 5 personas por lo menos dos de ellas hayan nacido el mismo día de la semana (la solución similar al del problema 23).

Rp. $1-V_5^7/7^5$

25. Se selecciona al azar un número de los números, $0,1,2,3,...,999$, calcular la probabilidad de que el número no sea divisible ni por tres ni por siete.

Rp. 0.572.

26. En el control de calidad de un artículo la probabilidad de que se encuentren por lo menos ocho artículos defectuosos es 0.15 y de que se encuentren a lo más 4 artículos defectuosos es 0.50, ¿cuál es la probabilidad de que se encuentren 5, 6, 7 artículos defectuosos en el control?. Rp. 0.35.

27. Una urna contiene 3 bolas numeradas de 1 a 3. Las bolas se sacan al azar una a una sin reposición. Si se considera un éxito cuando la bola k sale en la extracción k, $k=1,2,3$,

a) describa el espacio muestral
b) calcule la probabilidad de obtener al menos un éxito.

Rp. a) Ω={EEE, EFF, FFE, FFF, FFF, FEF}, b) 2/3.

28. Se va a seleccionar a 3 alumnos de 10 alumnos candidatos compuesto de 7 hombres y 3 mujeres para una determinada tarea. El seleccionador no sabe que los 10 alumnos están calificados de 1 a 10 según su eficiencia en esa tarea. Calcular la probabilidad de que la terna contenga

a) uno de los 2 mejores y dos de los 3 peores candidatos.
b) Por lo menos una mujer.

Rp. a) $C_1^2 C_2^8 / C_3^{10}$, b) $\sum_{i=1}^{3} C_{3-i}^7 C_i^3 \Big/ C_3^{10} = 1 - C_3^7 C_0^8 / C_3^{10}$

29. Los valores de 5 acciones diferentes en la bolsa son: 1,2,3,4,5 dólares respectivamente. Se eligen 3 acciones al azar de una en una sin restitución. Hallar la probabilidad de que la acción de mayor precio obtenida exceda la acción de menor precio por lo menos en 3$

Rp. CP=5×4×3=60, CF=42, mínimo 1 y máximo 4 ó 5 son 30 casos, mínimo 2 y máximo 5 son 12 casos, prob. = 42/60 = 0.7

30. Una empresa ha hecho un pedido de 20 "discos duros" que deben ser enviadas a tres de sus sucursales A, B y C. de manera que A reciba 8, B reciba 7 y C reciba 5. Se sabe que 3 de tales discos no han sido "formateados". Si la asignación se hace en forma aleatroria, ¿qué probabilidad hay de que los tres discos no formateados sean enviados a la misma sucursal?.

Rp. $(C_5^{17} C_7^{12} + C_4^{17} C_8^{13} + C_2^{17} C_8^{15}) / C_8^{20} C_7^{12}$

31. Los resultados de un experimento aleatorio caen en el intervalo [a, b], $a < b$, para cada intervalo A contenido en [a, b], definimos la probabilidad de A por:

$$P(A) = \text{longitud de } A/(b-a),$$

pruebe que P es una probabilidad.

32. El precio de un bien A puede tomar cualquier valor entre 0 y 6 soles, mientras que el precio de un bien B puede tomar cualquier valor entre 0 y 12 soles. Si usted está dispuesto a gastar más de 6 soles comprando tales bienes, calcule la probabilidad de que pueda comprar de 2 unidades de A y 3 unidades de B.

Rp. $\Omega=\{(x,y)/0\leq x\leq 6,\ 0\leq y\leq 12\}$, a) $A=\{(x,y)\in\Omega)/\ 2x+3y>6\}$, $P(A)=1-[(2\times 3)/2]=2/3$.

33. Un segmento de longitud 10 se divide en 3 segmentos. Calcular la probabilidad de que con tales segmentos se pueda construir un triángulo, sabiendo que la suma de las longitudes de dos lados es mayor o igual que la longitud del tercero.

Rp. 1/8.

34. Dos gerentes deciden encontrarse en cierto lugar para cerrar un negocio entre las 8 p.m y las 9 p.m de una día determinado. Si convienen que cada uno de ellos debe esperar al otro a lo más 10 minutos, ¿cuál es la probabilidad de que se encuentren?

Rp. $A=\{(x,y)\in\Omega/\ |x-y|\leq 10\}$, $P(A)=11/36$.

35. La demanda de cada uno de dos artículos A y B varía aleatoria e indistintamente en un rango de 2000 a 10000 toneladas. Se acuerda de bajar el precio de venta de ambos artículos, si la suma de sus demandas supera las 5000 toneladas, pero no es mayor de 7000 toneladas. ¿Cuál sería la probabilidad de que el precio de venta de ambos artículos baje?

Rp. $\Omega=\{(x,y)/2\leq x\leq 10,\ 2\leq y\leq 10\}$, $A=\{(x,y)\in\Omega)/5\leq x+y\leq 7\}$, $P(A)=4/64=1/16$

36. Un sistema está formado por dos componentes electrónicas cuyas duraciones varían aleatoriamente e indistintamente entre 0 y 5 años, pero la segunda componente actúa como respaldo de la primera tan solo cuando la primera deja de funcionar. Si la primera componte tiene mayor duración, calcular la probabilidad de que ésta dure dos o más años que la segunda.

Rp. Ω=triangulo: (0,0), (5,0), (5,5), $A=\{(x,y)\in\Omega)/x\geq y+2\}$, $P(A)=(3\times 3/2)/(5\times 5/2)=9/25$

37. Suponga que el precio P de un galón que un industrial fijará durante la semana siguiente para la venta de su solvente podría tomar indistintamente valores entre 10 soles y 20 soles. Un estudio ha revelado que la demanda semanal del solvente puede también tomar indistintamente valores entre $400-5P$ galones y los 1000 galones, ¿Con qué probabilidad el solvente tendrá durante la siguiente semana una demanda inferior a los 600 galones?

Rp. (255+ 250)/(255+250+4000)=2750/6750,.

5.5 Probabilidad condicional

A menudo se quiere determinar la probabilidad de que ocurra un evento sabiendo que otro evento ha ocurrido. La probabilidad *condicional* (o condicionada) de que un evento B ocurra dado que otro evento A ha ocurrido se denota por $P(B/A)$. Esta notación se lee "la probabilidad de que B ocurra dado que A ha ocurrido" o simplemente "la probabilidad de B dado A"

Definición. Sean A y B dos eventos en un espacio muestral. La probabilidad condicional de B dado A, es el número $P(B/A)$ que se define por:

$$P(B/A)=\frac{P(A\cap B)}{P(A)}, \text{ si } P(A)>0.$$

NOTAS.

1. Si $P(A)=0$, se define $P(B/A)=0$.
2. Observar que $(A\cap B)\subset A$, luego, cada vez que se calcula $P(B/A)$ estamos realmente calculando $P(B)$ con respecto al *espacio muestral reducido* A. Por esto, $P(B/A)$ se interpreta también como la *actualización* de $P(B)$ cuando el evento A ha ocurrido.
3. En particular, si A y B son dos eventos de un espacio muestral finito equiprobable Ω, la probabilidad *condicional* de B dado A, se calcula por

$$P(B/A)=\frac{\#(A\cap B)}{\#(A)}, \text{ si } \#(A)>0$$

4. Si $A\cap B=\varnothing$, entonces, $P(B/A)=0$
5. Si $A\subset B$, entonces, $P(B/A)=P(A/A)=1$
6. Si $B\subset A$, entonces, $P(B/A)=\dfrac{P(B)}{P(A)}$.

EJEMPLO 5.15.

Un club consiste de ciento cincuenta miembros. Del total, 3/5 son hombres y 2/3 son profesionales. Además, 1/3 de las mujeres son no profesionales.

a) Se elige al azar un socio del club:

a1) Calcular la probabilidad de que sea hombre y profesional.

a2) Calcular la probabilidad de que sea hombre, dado que es profesional.

b) Se eligen tres socios al azar:

b1) Si las tres son mujeres, ¿cuál es la probabilidad de que al menos una de ellas sea profesional?

b2) Si resultan ser del mismo sexo, ¿cuál es la probabilidad de que sean mujeres?.

SOLUCION.

El espacio muestral Ω consiste de los 150 miembros del club que son clasificados en: Hombre (*H*), Mujer(*M*), Profesional (*P*), y No profesional (*N*), según la tabla 5.1

Tabla 5.1

	Profesional (*P*)	No profesional (*N*)	Total
Hombre (*H*)	60	30	90
Mujer (*M*)	40	20	60
Total	100	50	150

a) Si se elige un socio al azar

a1) $P(H \cap P) = \dfrac{\#(H \cap P)}{\#(\Omega)} = \dfrac{60}{150} = 0.4$

a2) Considerando el **espacio muestral reducido** *P*, tenemos.

$$P(H/P) = \frac{\#(H \cap P)}{\#(P)} = \frac{60}{100} = 0.6 \cdot$$

También, usando la definición de probabilidad condicional, tenemos,

$$P(H/P) = \frac{P(H \cap P)}{P(P)} = \frac{60/150}{100/150} = 0.6 \cdot$$

b) Si se eligen tres socios al azar

b1) Sean los eventos A :" las tres son mujeres" y B :"al menos una es profesional" .

Considerando el espacio muestral reducido A se tiene,

$$P(B/A) = \frac{\#(A \cap B)}{\#(A)} = \frac{C_1^{40}C_2^{20} + C_2^{40}C_1^{20} + C_3^{40}C_0^{20}}{C_3^{60}} = 1 - \frac{C_0^{40}C_3^{20}}{C_3^{60}}$$

b2) Sean los eventos A :" los tres son del mismo sexo" y B :"las 3 son mujeres" . Observar que el evento A es "los tres son H o las tres son M", y que $B \subset A$, luego $B \cap A = B$. Entonces,

$$\#(A) = C_3^{90} + C_3^{60} = 151{,}700, \text{ y } \#(B) = C_3^{60} = 34{,}220$$

Considerando el espacio muestral reducido A, se tiene,

$$P(B/A)=\frac{\#(B)}{\#(A)}=\frac{34{,}220}{151{,}700}=0.23\cdot$$

NOTA. El lector debería verificar que para eventos en $\Omega : A$ (fijo) y B (cualquiera), la probabilidad condicional $P(B/A)$, satisface los diversos postulados de probabilidad. Esto es,

a) $P(B/A)$ satisface los axiomas de la probabilidad:

1) $P(B/A)\geq 0$ para todo evento B.

2) $P(\Omega/A)=1$

3) Si B_1 y B_2 son eventos que se excluyen mutuamente, entonces,

$$P((B_1\cup B_2)/A)=P(B_1/A)+P(B_2/A)$$

b) $P(B/A)$ satisface los teoremas de probabilidad, entre otros,

1) $P(\varnothing/A)=0$

2) $P(B^C/A)=1-P(B/A)$.

3) Si B_1 y B_2 son dos eventos cualesquiera, entonces,

$$P((B_1\cup B_2)/A)=P(B_1/A)+P(B_2/A)-P((B_1\cap B_2)/A)$$

5.6 Eventos independientes

Definición 1. Se dice que el evento B es *independiente* (estadísticamente o estocásticamente o vía probabilidad) del evento A, si,

$$P(B/A)=P(B).$$

Se verifica que, si $P(B/A)=P(B)$, entonces, $P(A/B)=P(A)$ y recíprocamente. En consecuencia, podemos afirmar que:

Los eventos A y B son independientes si, y sólo si,

$$P(B/A)=P(B) \text{ y } P(A/B)=P(A).$$

Esto equivale a decir que A y B son eventos independientes si y sólo si

$$P(A\cap B)=P(A)P(B).$$

Este resultado es con frecuencia utilizado como la definición de dos eventos independientes por su aplicación práctica. Establecemos la siguiente

Definición 2. Dos eventos A y B son independientes si y sólo si

$$P(A \cap B) = P(A)P(B).$$

De otra forma A y B **son dependientes.**

NOTA. A y B independientes no significa $A \cap B = \varnothing$.

EJEMPLO 5.16

Un dado se lanza 2 veces,

a) Si A es el evento "sale 2 en el primer lanzamiento", y si B es el evento "sale 5 en el segundo lanzamiento", son A y B independientes?.

b) Calcular la probabilidad de obtener los números 2 y 5.

SOLUCION.

El espacio muestral del experimento de lanzar un dado 2 veces es el conjunto

$$\Omega = \{(i, j) / i, j = 1,2,3,4,5,6\}.$$

a) El evento A:"sale 2 en el primer lanzamiento" es el conjunto:

$$A = \{(2, j) / j = 1,2,3,4,5,6\}$$

El evento B: "sale 5 en el segundo lanzamiento" es,

$$B = \{(i,5) / i = 1,2,3,4,5,6\}.$$

Además, el evento $A \cap B = \{(2,5)\}$.

Entonces,

$$P(A) = 1/6,\ P(B) = 1/6,\ P(A \cap B) = 1/36,\ P(A/B) = 1/6,\ \ P(B/A) = 1/6.$$

Por consiguiente, los eventos A y B son independientes ya sea por que,

$$P(B/A) = P(B) \quad \text{y} \quad P(A/B) = P(A),$$

o por que, $$P(A \cap B) = P(A)P(B).$$

b) Sean los eventos A_i : "sale 2 en la tirada i", $i = 1, 2$, y B_i: "sale 5 en la tirada i", $i = 1, 2$, entonces el evento C "sale los números 2 y 5 ", es,

$$C = A_1 \cap B_2 \cup B_1 \cap A_2.$$

Luego,

$$P(C) = P(A_1)P(B_2) + P(B_1)P(A_2) = \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} = \frac{2}{36}.$$

Observar también que,

$$CP = \#(\Omega) = 36, \quad A = \{(2,5),(5,2)\}, \quad CF = \#(A) = 2,$$

$$P(A) = \frac{CF}{CP} = \frac{2}{36}$$

TEOREMA 5.5.

Si los evento A y B son independientes, entonces:

a) A y B^C son independientes,

b) A^C y B son independientes,

c) A^C y B^C son independientes.

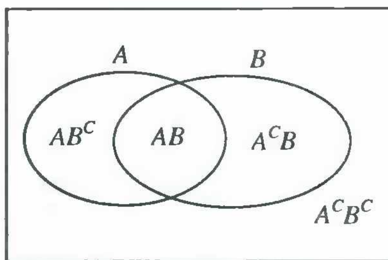


Figura 5.6

PRUEBA.

a) Se puede escribir el evento A como una unión disjunta (figura 5.6) de la forma $A = AB \cup AB^c$.

$$\text{Luego,} \quad P(A) = P(AB) + P(AB^c).$$

Además, por hipótesis, $P(AB) = P(A)P(B)$. En consecuencia,

$$P(AB^c) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c)\cdot$$

Para probar b), use la identidad: $B = AB \cup A^C B$.

Para probar c), use por ejemplo, la identidad: $\Omega = B \cup AB^C \cup A^C B^C$

EJEMPLO 5.17

Suponga que en un proceso de producción se utilizan las máquinas, 1 y 2, que trabajan en forma independiente para producir cierto bien. Si la probabilidad de que ambas máquinas fallen es 1/5 y de que falle sólo la 2 es 2/15. Calcular la probabilidad de que

a) falle sólo la máquina 1.

b) la producción continúe.

SOLUCION.

Sean los eventos: A: "falla la máquina 1", y B: "falla la máquina 2", entonces, $P(A\cap B)=1/5=3/15$, $P(A^c\cap B)=2/15$, entonces, $P(B)=5/15$

Además de $P(A\cap B)=P(A)P(B)$, resulta, $P(A)=9/15$

a) $P(AB^c)=P(A)P(B^c)=\dfrac{9}{15}\times\dfrac{10}{15}=\dfrac{6}{15}$

b) $P(A^cB\cup AB^c\cup A^cB^c)=P(A^cB)+P(AB^c)+P(A^cB^c)$

$$=P(A^c)P(B)+P(A)P(B^c)+P(A^c)P(B^c)$$

$$=\frac{6}{15}\times\frac{5}{15}+\frac{9}{15}\times\frac{10}{15}+\frac{6}{15}\times\frac{10}{15}=\frac{12}{15}.$$

Observar también que,

P(prod continue)=1−P(prod no continue)=$1-P(AB)=1-P(A)P(B)=12/15$.

Definición. Diremos que tres eventos A, B y C son independientes si, y sólo si, verifican las relaciones:

$$P(A\cap B)=P(A)P(B)$$

$$P(A\cap C)=P(A)P(C),$$

$$P(B\cap C)=P(B)P(C)$$

$$P(A\cap B\cap C)=P(A)P(B)P(C).$$

Observar que si, sólo se verifica: $P(A\cap B\cap C)=P(A)P(B)P(C)$ no se puede concluir que los eventos A, B, C sean independientes.

En general, se dice que la sucesión finita de n eventos $A_1,A_2,\cdots,A_n$, son **independientes** si, y sólo si, la probabilidad de la intersección de dos, de tres, ..., de $n-1$ cualesquiera y de los n eventos, es igual al producto de las probabilidades de los dos, de los tres, ..., de los $n-1$ y de los n. El total de relaciones que se verifican es:

$$C^n_2+C^n_3+\ldots+C^n_n=2^n-C^n_0-C^n_1=2^n-n-1.$$

5.7 Reglas de la multiplicación

TEOREMA 5.6

a) $P(A \cap B) = P(A)P(B)$, si A y B son dos eventos independientes

$P(A \cap B) = P(A)P(B/A)$, si B depende de A.

(La segunda proviene de la definición de probabilidad condicionada)

b) En general, si $A_1, A_2, \cdots, A_n$ son n eventos, entonces,

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n)$$

si los eventos son independientes.

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2)\ldots P(A_n/A_1 \cap A_2 \cap \ldots \cap A_{n-1}).$$

si los eventos son dependientes, y siempre que $P(A_1 \cap A_2 \cap \ldots A_{n-1}) > 0$.

La pruebas (por inducción) se dejan como ejercicios para el lector.

NOTAS.

1) Si los eventos son independientes, entonces,

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n)$$

Por contraposición en b), si,

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) \neq P(A_1)P(A_2)\ldots P(A_n).$$

entonces los n eventos $A_1, A_2, \cdots, A_n$ no son independientes.

2) Si los n eventos $A_1, A_2, \cdots, A_n$ son independientes, entonces,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - P\left(\bigcap_{i=1}^{n} (A_i)^c\right) = 1 - \prod_{i=1}^{n} P(A_i^c)$$

EJEMPLO 5.18

Un lote contiene 15 objetos de los cuales 7 son calificados como E (éxito) y el resto como F (fracasos). Del lote se escogen 5 objetos al azar una tras otra, calcular la probabilidad de que los cinco sean éxitos, si las extracciones se hacen:

a) con reposición,
b) sin reposición.

SOLUCION.

Sean los eventos:

E_i : "sale objeto defectuoso en la extracción i", i = 1,2,3,4,5.

F_i : "sale objeto bueno en la extracción i", i = 1,2,3,4,5.

Si A es el evento "los 5 objetos sustraídos sucesivamente son éxitos", entonces,

$$A = E_1 \cap E_2 \cap E_3 \cap E_4 \cap E_5 = E_1 E_2 E_3 E_4 E_5$$

a) Si las extracciones son con reposición **los eventos son independientes**, entonces,

$$P(A) = P(E_1)P(E_2)P(E_3)P(E_4) = \frac{7}{15} \times \frac{7}{15} \times \frac{7}{15} \times \frac{7}{15} \times \frac{7}{15}$$

b) Si las extracciones son sin reposición **los eventos son dependientes**, entonces,

$$P(A) = P(E_1)P(E_2/E_1)P(E_3/E_1E_2)P(E_4/E_1E_2\,E_3)P(E_5/E_1E_2\,E_3E_4)$$

$$P(A) = \frac{7}{15} \times \frac{6}{14} \times \frac{5}{13} \times \frac{4}{12} \times \frac{3}{11}.$$

Observar que en este caso:

$$P(A) = \frac{7 \times 6 \times 5 \times 4 \times 3}{15 \times 14 \times 13 \times 12 \times 11} = \frac{V_5^7}{V_5^{15}}.$$

EJEMPLO 5.19

Con referencia al ejemplo 5.18, si se escogen 5 objetos al azar, calcular la probabilidad de que tres sean éxitos (y dos sean fracasos), si se escogen,

a) uno por uno con reposición,

b) uno por uno sin reposición,

SOLUCION.

Sea el evento, C: "tres E y dos F".

Un evento del evento compuesto C es $EEEFF$ cuya probabilidad es:

Caso a) $$P(EEEFF) = \frac{7}{15} \times \frac{7}{15} \times \frac{7}{15} \times \frac{8}{15} \times \frac{8}{15} = p^3 q^2,$$

donde $p = 7/15$, $q = 8/15$.

Caso b) $$P(EEEFF) = \frac{7}{15} \times \frac{6}{14} \times \frac{5}{13} \times \frac{8}{12} \times \frac{7}{11} = \frac{V_3^7 V_2^8}{V_5^{15}}$$

Todos los eventos de C, se obtienen ordenando $EEEFF$ de $C_3^5 = 10$ formas. Las probabilidades de cada una de estas formas son iguales a p^3q^2 en el caso a), y a $V_3^7 V_2^8 / V_5^{15}$ en el caso b). Por tanto, la probabilidad de C es:

Caso a) $\quad P(C) = C_3^5 p^3 q^2$, donde $p = 7/15$, $q = 8/15$.

Caso b) $\quad P(C) = C_3^5 \dfrac{V_3^7 V_2^8}{V_5^{15}}$

NOTA. Si los cinco objetos se escogen a la vez, entonces, $P(C) = \dfrac{C_3^7 C_2^8}{C_5^{15}}$.

El lector debería verificar que : $\dfrac{C_3^5 V_3^7 V_2^8}{V_5^{15}} = \dfrac{C_3^7 C_2^8}{C_5^{15}}$.

NOTA. En general, la probabilidad de obtener k éxitos (E) ($n-k$ fracasos (F)) al escoger n objetos uno por uno sin reposición de un lote de N objetos de los cuales r son éxitos y el resto fracasos es igual al número:

$$C_k^n V_k^r V_{n-k}^{N-r} \Big/ V_n^N = C_k^r C_{n-k}^{N-r} \Big/ C_n^N .$$

EJEMPLO 5.20

Un experimento aleatorio consiste en la realización consecutiva de pruebas independientes, donde el resultado de cada prueba es un éxito (E) con probabilidad p, o un no éxito (E^c) con probabilidad $1-p$.

a) Calcular la probabilidad de que ocurra al menos un éxito en n pruebas ($n \geq 1$).
b) Si $p = 0.99$, ¿cuántas pruebas deberían realizarse para tener al menos un éxito con probabilidad 0.9999?.
c) ¿Cuál es la probabilidad de que el primer éxito en la k-ésima prueba ($k \geq 1$)?
d) ¿Cuál es la probabilidad de que ocurran k éxitos n pruebas ($k \leq n$)?.
e) ¿Cuál es la probabilidad de que ocurran k éxitos en n pruebas, de manera que ocurra el k-ésimo éxito en la n-ésima prueba ($k \leq n$)?

SOLUCION.

Sean los eventos:

E_i : "ocurre éxito en la prueba i", $i = 1, 2, \ldots, n$.

E_i^c : "no ocurre éxito en la prueba i", $i = 1, 2, \ldots, n$.

a) Si A es el evento "ocurre al menos un éxito en las n pruebas", entonces,

$$A=\bigcup_{i=1}^{n}E_i,\ \text{ y }\ P(A)=P\left[\bigcup_{i=1}^{n}E_i\right]=1-P\left[\left(\bigcup_{i=1}^{n}E_i\right)^c\right]=1-P\left[\bigcap_{i=1}^{n}E_i^c\right]$$

$$=1-\prod_{i=1}^{n}P\left(E_i^c\right)=1-\prod_{i=1}^{n}(1-p)=1-(1-p)^n$$

Luego, $$P(A)=1-(1-p)^n\ .$$

b) $P(A)=1-(1-p)^n=0.9999$, entonces, $(0.01)^n=0.0001$, $n=2$.

c) $P(\text{1er éxito en prueba } k)=q^{k-1}p$

d) Si B es el evento: "ocurren k éxitos en n pruebas", entonces, B consiste de C_k^n eventos que contienen cada uno, k éxitos (E) y $n-k$ no éxitos (E^c), luego,

$$P(B)=C_k^n\,p^k(1-p)^{n-k}.$$

Esta probabilidad se denomina probabilidad binomial

e) Sea el evento D: "ocurren k éxitos en n pruebas, de manera que el k-ésimo éxito sea la n-ésima prueba ". El último es un éxito, en las restantes ocurre una binomial. Luego,

$$P(D)=C_{k-1}^{n-1}p^k q^{n-k}$$

EJEMPLO 5.21.

Para decidir si se acepta o no un lote de 20 artículos en donde existen 4 defectuosos, se toman dos artículos al azar y a la vez. Si los dos son defectuosos se rechaza el lote, si los dos son buenos se acepta el lote y si solamente uno es bueno se toman otros dos artículos al azar a la vez de los 18 que quedan. Esta vez, si alguno es bueno se acepta el lote, de otro modo se rechaza. Calcular la probabilidad de aceptar el lote.

SOLUCION

P(aceptar el lote) = P(BB) + P(BD)×P[(BD o BB) /BD]

$$P(\text{aceptar el lote})=\frac{C_2^{16}}{C_2^{20}}+\frac{C_1^4C_1^{16}}{C_2^{20}}\left(\frac{C_1^3C_1^{15}}{C_2^{18}}+\frac{C_2^{15}}{C_2^{18}}\right)=\frac{932}{969}=0.962$$

Observar que: P[Aceptar] = 1 − P[Rechazar]

$$=1-\frac{C_2^4}{C_2^{20}}+\frac{C_1^4C_1^{16}}{C_2^{20}}\times\frac{C_2^3C_0^{15}}{C_2^{18}}=1-\frac{37}{969}$$

EJEMPLO 5.22

Un lote contiene 40 artículos de los cuales 4 son defectuosos y 36 son no defectuosos. Si se divide al azar el lote en cuatro sublotes de 10 artículos cada uno, calcular la probabilidad de que en cada sublote haya un artículo defectuoso.

SOLUCION.

Sean los eventos A_i :"El lote i contiene un defectuoso", $i = 1,2,3,4$

y sea A: "Los cuatro sublotes contienen un artículo defectuoso cada uno", entonces, $A = A_1A_2A_3A_4$ y

$$P(A) = P(A_1A_2A_3A_4) = P(A_1)P(A_1/A_2)P(A_3/A_1A_2)P(A_4/A_1A_2A_3)$$

$$P(A) = \frac{C_1^4C_9^{36}}{C_{10}^{40}} \times \frac{C_1^3C_9^{27}}{C_{10}^{30}} \times \frac{C_1^2C_9^{18}}{C_{10}^{20}} \times \frac{C_1^1C_9^{9}}{C_{10}^{10}} = \frac{4!\times 36!/9!^4}{40!/10!^4}.$$

También observar que:

$$CP = P_{10,10,10,10}^{40} = \frac{40!}{10!^4}, \quad CF = 4!\,P_{9,9,9,9}^{36} = \frac{4!\times 36!}{9!^4}$$

$$P(A) = \frac{CF}{CP} = \frac{4!\times 36!/9!^4}{40!/10!^4}.$$

5.8 Regla de la probabilidad total y regla de Bayes

5.8.1 Partición del espacio muestral

Definición. Se denomina **partición** del espacio muestral Ω, a una colección de k eventos $A_1, A_2, ..., A_k$ que sean mutuamente excluyentes y cuya unión es el espacio muestral Ω, es decir, tales que verifican las siguientes condiciones:

1) $P(A_i) > 0$, para cada $i = 1,2,...,k$

2) $A_i \cap A_j = \varnothing \quad \forall i \neq j$

3) $\bigcup_{i=1}^{k} A_i = \Omega$. o $\sum_{i=1}^{k} P(A_i) = 1$

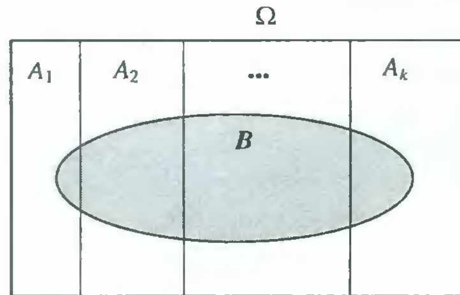


Figura 5.7. Partición de Ω

5.8.2 Regla de la probabilidad total y regla de Bayes

TEOREMA 5.7. (Regla de la probabilidad total).

Si k eventos: $A_1, A_2, ..., A_k$, constituyen una partición del espacio muestral Ω, entonces, para cualquier evento B en Ω,

$$P(B) = \sum_{i=1}^{k} P(A_i)P(B/A_i).$$

PRUEBA.

Para cualquier evento B de Ω (figura 5.7)

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{k} A_i \right) = \bigcup_{i=1}^{k} (B \cap A_i).$$

Además, los eventos $B \cap A_i$, $i = 1, 2, ..., k$, son excluyentes dos a dos. Luego, por el axioma P3), se tiene:

$$P(B) = \sum_{i=1}^{k} P(A_i \cap B) = \sum_{i=1}^{k} P(A_i)P(B/A_i)$$

TEOREMA 5.8. (Regla de Bayes).

Si los *k* eventos: $A_1, A_2, ..., A_k$, constituyen una partición del espacio muestral Ω, entonces, para cualquier evento *B* de Ω tal que $P(B) > 0$,

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{P(B)}, \text{ para cada } i = 1, 2, \cdots k$$

PRUEBA.

Para cualquier evento A_i de la partición, se tiene:

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B/A_i)}{P(B)}.$$

donde
$$P(B) = \sum_{i=1}^{k} P(A_i)P(B/A_i) \cdot$$

La regla de Bayes nos permite comparar la probabilidad *previa* (o *apriori*) $P(A_i)$ con la probabilidad *posterior* (o *aposteriori*) $P(A_i/B)$. Esto es, $P(A_i/B)$ es la probabilidad de A_i **corregida** o **modificada** por la ocurrencia del evento *B*.

Observar que la regla de Bayes *da el porcentaje de la contribución* de $P(A_i \cap B)$ con respecto a *P*(*B*).

EJEMPLO 5.23

Un ensamblador de computadoras usa partes que provienen de tres proveedores P_1, P_2 y P_3. De 2000 partes recibidas 1000 provienen de P_1, 600 de P_2 y el resto de P_3. De experiencias pasadas, el ensamblador sabe que las partes defectuosas que provienen de P_1, P_2 y P_3 son respectivamente 3%, 4%, y 5%. Si se elige una computadora al azar,

a) ¿Cuál es la probabilidad de que contenga una parte defectuosa?

b) y si contiene una parte defectuosa, ¿cuál es la probabilidad de que haya sido proveído por P_2?.

SOLUCION.

Sean los eventos:

A_i: "Parte proviene del proveedor P_i", $i = 1,2,3$.

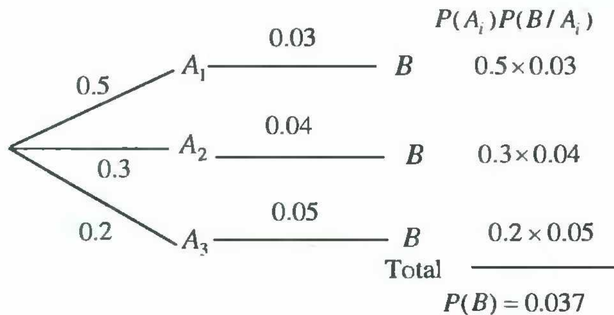
B: "Parte defectuosa".

Las probabilidades, *P*(A_i), *P*(*B*/A_i) se ubican en el diagrama que sigue:

donde

$$P(A_1)=\frac{1000}{2000}=0.5\,,\quad P(A_2)=\frac{600}{2000}=0.3\,,\quad P(A_3)=\frac{400}{2000}=0.2$$

$$P(B/A_1)=0.03\,,\quad P(B/A_2)=0.04\,,\quad P(B/A_3)=0.05$$



a) Aplicando la regla de la probabilidad total, se obtiene,

$$P(B)=P(A_1)P(B/A_1)+P(A_2)P(B/A_2)+P(A_3)P(B/A_3)$$

$$P(B)=0.5\times 0.03+0.3\times 0.04+0.2\times 0.05=0.037\,.$$

b) Aplicando la regla de Bayes, se obtiene,

$$P(A_2/B)=\frac{P(A_2)P(B/A_2)}{P(B)}=\frac{0.3\times 0.04}{0.037}=0.3243\cdot$$

EJEMPLO 5.24

En un día cualquiera cuatro máquinas M_1, M_2, M_3 y M_4 producen un bien de consumo en las siguientes proporciones: M_1 produce el doble de M_4, M_3 produce el triple de M_4, mientras que M_1 produce la mitad de M_2. Las producciones *no* defectuosas son respectivamente 95%, 95%, 90% para M_1, M_2, M_3. Si se elige al azar un artículo de la producción de un día y se encuentra que la probabilidad de que resulte no defectuoso es 0.93

a) ¿Cuál es el porcentaje de producción no defectuosa de M_4?.

b) ¿De qué máquina es más probable que provenga un artículo defectuoso?.

SOLUCION.

Sean los eventos:

A_i: "el artículo extraído es producido por la máquina M_i ", $i = 1,2,3,4$.
B : "el artículo extraído es no defectuoso".

Si $\#(A_4) = x$, entonces, $\#(A_1) = 2x$, $\#(A_3) = 3x$, $\#(A_2) = 4x$,

Además, $\#(\Omega) = 10x$.
Por lo tanto,

$$P(A_1) = 2/10,\ P(A_2) = 4/10,\ P(A_3) = 3/10,\ P(A_4) = 1/10$$

$$P(B/A_1) = 0.95,\ P(B/A_2) = 0.95,\ P(B/A_3) = 0.90$$

a) Aplicando la regla de la probabilidad total, se tiene.

$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3) + P(A_4)P(B/A_4)$$

$$P(B) = 0.2 \times 0.95 + 0.4 \times 0.95 + 0.3 \times 0.90 + 0.1 \times P(B/A_4) = 0.93$$

Luego, $P(B/A_4) = 0.90$

b) La probabilidad de producción defectuosa es:

$$P(B^c) = 0.2 \times 0.05 + 0.4 \times 0.05 + 0.3 \times 0.10 + 0.1 \times 0.10 = 0.0.7$$

Aplicando la regla de Bayes para cada A_i, $i = 1,2,3,4$, resulta que $P(A_3/B^c) = 0.03/0.07$ es mayor que el resto. Por lo tanto, si el artículo escogido al azar es defectuoso, es más probable que provenga de la producción de la máquina M_3.

EJERCICIOS RESUELTOS

EJERCICIO 1

Las probabilidades de que los socios S_1 y S_2 sean elegidos presidente de su club son respectivamente 0.4 y 0.6. Las probabilidades de que se aumenten las cuotas mensuales de los socios son de 0.9 si sale elegido S_1 y de 0.2 si sale elegido S_2,

a) ¿Cuál es la probabilidad de que haya un aumento en las cuotas mensuales de los socios?.
b) Si se aumenta la cuota mensual, ¿cómo se modifican las probabilidades de que salgan elegidos los socios S_1 y S_2?.

SOLUCION.

Sean los eventos:

A_i : "Sale elegido socio S_i ", i =1, 2, y

B: "Se incrementan las cuotas mensuales de los socios".

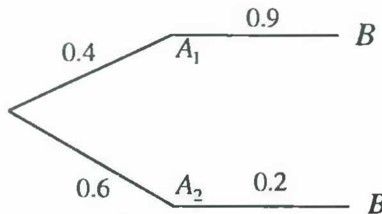
Las probabilidades:

$$P(A_1)=0.4,\ P(A_2)=0.6,\ P(B/A_1)=0.9,\ P(B/A_2)=0.2$$

se representan en el diagrama de árbol que sigue:

Por tanto,

a) $P(B)=P(A_1)P(B/A_1)+P(A_2)P(B/A_2)=0.4\times0.9+0.6\times0.2=0.48.$



b) $P(A_1/B)=\dfrac{P(A_1)P(B/A_1)}{P(B)}=\dfrac{0.4\times0.9}{0.48}=0.75\cdot$

$$P(A_2/B)=\frac{P(A_2)P(B/A_2)}{P(B)}=\frac{(0.6)(0.2)}{0.48}=0.25\cdot$$

La probabilidad de A_1 se modifica de 0.4 a 0.75 y la de A_2 de 0.6 a 0.25.

En consecuencia, se puede concluir que, si se aumentan las cuotas mensuales, probablemente el socio S_2 no sea elegido presidente de su club.

EJERCICIO 2.

Se estima que la probabilidad de que una compañía B tenga éxito al comercializar un producto es de 0.95 si su competidora la compañía A no interviene en el mercado, y es de 0.15 si la Cía. A interviene en el mercado. Si se estima que A intervendría en el mercado con probabilidad de 0.7 .

a) ¿Cuál es la probabilidad de que la Cía. B tenga éxito?

b) Si la compañía B no tuviera éxito, ¿en cuanto se estima la probabilidad de que A intervenga en el mercado?

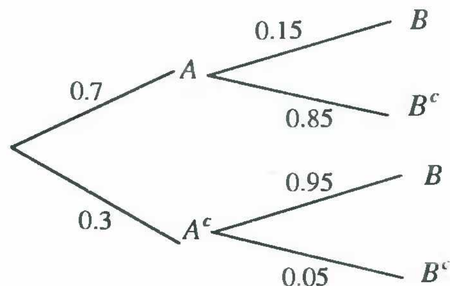
SOLUCION.

Sean los eventos B: "la Cía. B tiene éxito" y A: "la Cía. A interviene en el mercado". Las probabilidades:

$$P(A)=0.7\,,\; P(A^c)=0.3\,,\; P(B/A)=0.15\,,\; P(B/A^c)=0.95\,,$$
$$P(B^c/A)=0.85\,,\; P(B^c/A^c)=0.05$$

se representan en el siguiente diagrama de árbol.
Por tanto,

a) $P(B)=P(A)P(B/A)+P(A^c)P(B/A^c)=0.7\times0.15+0.3\times0.95=0.39$.



b) $P(B^c)=P(A)P(B^c/A)+P(A^c)P(B^c/A^c)=0.7\times0.85+0.3\times0.05=0.61$

Si B no tuviera éxito, la probabilidad de que A intervenga en el mercado es:

$$P(A/B^c)=\frac{P(A)P(B^c/A)}{P(B^c)}=\frac{0.7\times0.85}{0.61}=0.975\cdot$$

EJERCICIO 3.

En un proceso de producción se sabe que durante cuatro décimas partes de tiempo se producen 20% de unidades defectuosas y durante seis décimas partes de tiempo se producen 15% de unidades defectuosas. De la producción que consiste de 20 unidades de sólo una de las modalidades, se inspeccionan tres elegidos al azar a la vez y se encuentran dos unidades defectuosa. En base a este resultado, ¿qué modificaciones acerca de las probabilidades de las dos calidades de producción se deben hacer?.

SOLUCION.

Sean los eventos

A_1: "calidad del 20% defectuoso",

A_2 : "calidad del 15% defectuoso", y

B : "dos defectuosos en la muestra de tres",

Entonces: $P(A_1) = 0.4$, $P(A_2) = 0.6$,

$$P(B/A_1) = C_2^4 C_1^{16} \big/ C_3^{20} \quad \text{y} \quad P(B/A_2) = C_2^3 C_1^{17} \big/ C_3^{20}$$

$$P(B) = 0.4\, C_2^4 C_1^{16} \big/ C_3^{20} + 0.6\, C_2^3 C_1^{17} \big/ C_3^{20} = 69/1140 \cdot$$

$$P(A_1/B) = 38.4/69 = 0.56 \quad \text{y} \quad P(A_2/B) = 30.6/69 = 0.44 \cdot$$

La calidad del 20% defectuoso tiene probabilidad corregida igual a 0.56, mientras que la calidad 15% defectuoso tiene la probabilidad corregida igual a 0.44.

EJERCICIO 4.

Un lote que contiene 12 artículos de los cuales x son defectuosos y el resto no defectuosos es sometido a dos controles. En el primer control se extrae de este lote un artículo al azar , si está bueno se le devuelve al lote y si está defectuoso se lo reemplaza por uno bueno, luego se pasa el lote al segundo control.

a) Determinar el número artículos defectuosos que hay en el lote si este pasa al segundo control de forma tal que la probabilidad de extraer al azar aquí un artículo *no* defectuoso es 61/72.

b) Suponga que el lote pasa al segundo control con el número de defectuosos hallados en la parte a). El segundo control consiste en extraer 3 artículos al azar a la vez y rechazar el lote si se encuentran al menos dos artículos defectuosos, calcular la probabilidad de aceptar el lote.

SOLUCION.

a) Sean los eventos,

D_i : "sale artículo defectuoso en el control i", $i = 1,2$.

B_i : "sale artículo bueno en el control i", $i = 1,2$.

La probabilidad de que el artículo sea *no* defectuoso en el segundo control es igual a $P(B_2) = 61/72$. Además,

$$P(B_2) = P(D_1)P(B_2/D_1) + P(B_1)P(B_2/B_1)$$

$$P(B_2) = \left(\frac{x}{12}\right)\left(\frac{13-x}{12}\right) + \left(\frac{12-x}{12}\right)\left(\frac{12-x}{12}\right) = \frac{11x+144}{(12)^2}.$$

Entonces, $\dfrac{-11x+144}{(12)^2} = \dfrac{61}{72}$, de donde resulta $x = 2$.

b) El lote es aceptado, si en la terna de artículos no hay defectuosos o hay un sólo defectuoso.

Sean los eventos A_i:"los tres artículos extraídos contienen i defectuosos"

$i = 0, 1$. La probabilidad de aceptar el lote es:

$$P(A_0 \cup A_1) = P(A_0) + P(A_1) = \frac{C_0^2 C_3^{10}}{C_3^{12}} + \frac{C_1^2 C_2^{10}}{C_3^{12}} = \frac{21}{22}.$$

EJERCICIO 5.

Un vendedor tiene una lista de 5 clientes a quienes frecuentemente les ofrece un producto. La probabilidad de que uno de tales clientes le compre el producto es distinta para cada cliente y está dada por $P_i = (i-1)/4$, para $i = 1, 2, 3, 4, 5$.

a) Si selecciona uno de los clientes al azar, ¿cuál es la probabilidad de que este le compre el producto?

b) Si selecciona uno de los clientes al azar y este le compra el producto una primera vez, ¿cuál es la probabilidad de que en una segunda selección al azar este mismo cliente le compre el producto?.

SOLUCION.

Sean los eventos:

A_i: "El vendedor escoge al i–esimo cliente ", $i = 1, 2, 3, 4, 5$.

B: "El cliente le compra el producto" .

$$P(A_i) = 1/5 \text{ y } P(B/A_i) = (i-1)/4 \text{ para: } i = 1, 2, 3, 4, 5$$

Entonces,

a) $P(B) = \sum_{i=1}^{5} P(A_i)P(B/A_i) = (1/5)(0/4 + 1/4 + 2/4 + 3/4 + 4/4) = 1/2$

b) Si el cliente le compra el producto una primera vez, la probabilidad de que haya escogido al i-ésimo cliente es $P(A_i / B) = (1/5)(i-1)/1/2 = (i-1)/10$

Sea C: "El mismo cliente seleccionado una segunda vez le compra el producto"
Entonces, $P(C/(A_i/B)) = (i-1)/4$ para $i = 1, 2, 3, 4, 5$

Luego,

$$P(C) = \sum_{i=1}^{5} P(A_i/B)P(C/(A_i/B)) = (0/10)(0/4) + (1/10)(1/4) + (2/10)(2/4) + (3/10)(3/4) + (4/10)(4/4) = 3/4$$

EJERCICIOS

1. si $P(A)=5/8$, $P(B)=3/4$ y $P(A/B)=2/3$, calcular $P(A/B^c)$.

Rp. 1/2.

2. Si $P(B)=3/15$, $P(B/A)=1/5$, y $P(A\cap B)=1/15$, calcular $P(A\cap B^c)$.

Rp. 4/15.

3. En una muestra de 120 loretanos se encontró que el 60% sufre alguna enfermedad, el 30% tienen al menos 30 años, y el 20% son menores de 30 años y sanos. Si uno de tales loretanos es escogido al azar, ¿cuál es la probabilidad
a) de que sufra alguna enfermedad y tenga al menos 30 años?.
b) de que sufra alguna enfermedad si tiene al menos 30 años?.

Rp. a) 12/120, b) 12/36.

4. De 200 clientes de crédito de una tienda comercial, 100 tienen créditos menores que $200, 15 tienen créditos de al menos $500, y 110 tienen créditos menores de 4 años. Además 30 clientes tienen créditos de al menos 4 años y de 200 a menos de $500, y 10 clientes tienen créditos de al menos $500 y menos de 4 años.
a) Si se elige un cliente al azar, ¿cuál es la probabilidad de que tenga crédito menos de 4 años si tiene saldo de crédito de menos de $200?.
b) Si se eligen dos clientes al azar y resultan de al menos de 4 años de crédito, ¿cuál es la probabilidad de que uno tenga saldo de crédito de $500 o más?.

Rp. a) 45/100, b) $C_1^5 C_1^{85} / C_2^{90}$.

5. En una encuesta de opinión se encontró que el 25% de los electores votarían por el candidato E. De los que no votarían por E el 20% son mujeres y el resto son hombres. Además la probabilidad de que un elector elegido al azar sea hombre es 0.70. Si se elige un elector al azar y resulta ser mujer, ¿cuál es la probabilidad de que no vote por E?

Rp. 0.15/0.30.

6. Un comerciante recibe para su venta 80 objetos, 2/5 del proveedor A y el resto del proveedor B. El 12.5% de objetos de cada proveedor son defectuosos. Si se hace una inspección de cuatro objetos escogidos al azar a la vez y si resultan
a) ser de B, ¿cuál es la probabilidad de que al menos uno sea defectuoso?.
b) tres defectuosos, ¿cuál es la probabilidad de que dos de los defectuosos provengan de A?.

Rp. a) $1-(C_4^{42}/C_4^{48})$, b) $C_2^4 C_1^6 C_1^{70}/C_3^{10} C_1^{70} = C_2^4 C_1^6 / C_3^{10}$

7. En horas de trabajo, una cervecería utiliza dos máquinas embotelladoras M1 y M2, pero no operan simultáneamente. La probabilidad de que la primera

máquina se descomponga es 0.2. Si la primera máquina se descompone se enciende la segunda, la cual tiene probabilidad de descomponerse de 0.3. ¿Qué probabilidad hay de que el sistema embotellador no esté funcionando en las horas de trabajo?.

Rp. M_i: falla máquina i:1,2, $P(M_1 \cap M_2) = P(M_1)P(M_2/M_1) = 0.2 \times 0.3 = 0.06$

8. En un lote de 50 artículos, hay 10 de tipo A y 40 de tipo B, se extraen del lote 5 artículos al azar uno por uno si reposición, ¿cuál es la probabilidad de que al menos uno de estos sea de tipo A?.

Rp. P(al menos uno de tipo A)=1−P(ninguno de tipo A)=0.69

9. Sólo una de las 10 llaves que lleva una persona abre la cerradura de su puerta. El prueba las llaves una por una escogiendo al azar cada vez una de las llaves no probadas. Calcular la probabilidad de que la llave que abre la cerradura sea escogida en el quinto intento. Rp. 0.1

10. En una urna hay tres balotas numeradas de 1 a 3. Las balotas se sacan al azar una a una y sin reemplazo. Si la balota numerada con r se saca en la r-ésima extracción se considera un éxito. Hallar la probabilidad de obtener un éxito.

Rp. 1/2.

11. Se prueba un lote de 48 focos uno por uno (sin reposición). Si el lote contiene dos defectuosos, ¿cuál es la probabilidad de que el último defectuoso se detecte en la tercera prueba? Rp. $P(D_1B_2D_3) + P(B_1D_2D_3) = 0.0018$

12. La urna 1 contiene dos bolas rojas y dos bolas azules, mientras que la urna 2 contiene una bola roja y tres azules. Una bola es seleccionada aleatoriamente de la urna 1 y colocada en la urna 2. Luego una bola es seleccionada al azar de la urna 2 y colocada en la urna 1. Si ahora una bola es seleccionada al azar de la urna 1, ¿cuál es la probabilidad de que esta sea roja?. Rp. 9/20.

13. Una urna contiene 5 fichas rojas y algunas fichas blancas. Se extrae al azar una ficha de la urna y se reemplaza por una del otro tipo. Luego se saca de la urna una segunda ficha. Determinar el número de fichas blancas en la urna si se sabe que la probabilidad de que la segunda ficha sea roja es 0.5. Rp. 5.

14. Para decidir si se acepta o no un lote de 12 objetos en donde existen 3 defectuosos, se toman dos objetos al azar y a la vez. Si los dos son defectuosos, se rechaza el lote; si los dos son buenos se acepta el lote, y si sólo uno es bueno se toman otros dos objetos al azar y a la vez de los 10 que quedan. Esta vez, si alguno es bueno se acepta el lote, de otro modo se rechaza. Calcular la probabilidad de aceptar el lote.

Rp. 36/66 + (27/66)(44/45).

15. Si $P(A) = 1/3$ y $P(A \cup B) = 11/21$, calcular $P(B)$

a) si los eventos A y B son excluyentes.

b) si los eventos A y B son independientes. Rp. a) 4/21. b) 2/7.

16. Sea el espacio muestral: $\Omega = \{w_1, w_2, w_3, w_4\}$, donde,

$$P(\{w_1\}) = 1/4, \quad P(\{w_2\}) = 1/4, \quad P(\{w_3\}) = 1/4, \quad P(\{w_4\}) = 1/4$$

Sean los eventos: $A = \{w_1, w_2\}$, $B = \{w_1, w_3\}$, $C = \{w_1, w_4\}$. ¿Son los eventos A, B, y C independientes?.

Rp. no, pues no se verifica, $P(ABC)=P(A)P(B)P(C)$.

17. Pruebe que:

a) Si el evento B es independiente del evento A, entonces, A es independiente de B.

b) A y B son eventos independientes, si y sólo si,

$$P(A \cap B) = P(A)P(B)$$

c) Si A y B son eventos independientes, entonces, $P(B/A) = P(B/A^c)$

18. Un negocio es tal que su probabilidad de éxito es p. El negocio se realiza dos veces de manera independiente. ¿Qué valor de p hace máxima la probabilidad

a) de obtener éxito una sola vez?

b) de obtener éxito al menos una vez?

Rp. a) Prob=$2p(1-p)$, Prob. es máxima si $p=1/2$, b) $2p(1-p)+p^2$, es máximo, si $p=1$

19. Pruebe que todo evento de probabilidad cero o uno es independiente de cualquier otro evento.

Rp. Si $P(A)=0$, de $A\cap B\subset A$, $P(A\cap B)=0=P(A)P(B)$. Si $P(A)=1$, de $P(B)=P(A\cap B)+P(A^cB)$, $P(A\cap B)=P(B)=P(A)P(B)$, ya que $P(A^cB)\leq P(A^c)=0$.

20. Suponga que una compañía utiliza un procedimiento de prueba que es confiable en 98%. -Es decir identifica correctamente a un objeto como defectuosos o no defectuoso con una probabilidad de 0.98-. En un esfuerzo por reducir la probabilidad de error a cada objeto se somete a dos pruebas independientes

a) ¿Cuál es la probabilidad de que un objeto no defectuosos no pase ambas pruebas?

b) ¿Cuál es la probabilidad de que se detecte a un objeto defectuoso, es decir de que no pase por lo menos una de las dos pruebas?

Rp a) 0.02×0.02, b) $0.98\times0.02+0.02\times0.98+0.98\times0.98$

21. Una urna contiene 10 objetos numerados de 1 a 10. Un juego consiste en sacar tales objetos y termina cuando sale el numerado con uno. ¿Cuál es la probabilidad de que el juego termine si se sacan al azar 5 objetos

a) a la vez?.

b) uno a uno sin reposición?.

c) uno a uno con reposición?. Rp. a) 0.5, b) 0.1, c) $9^4/10^5$.

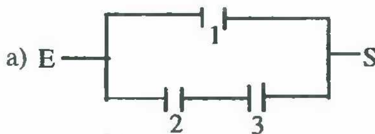
22. Se ha determinado que el porcentaje de televidentes que ven los programas A, B y C son respectivamente 0.4, 0.5, y 0.3. Cada televidente ve los programas independientemente uno del otro. Si se elige al azar a uno de tales televidentes, ¿qué probabilidad hay de que vea

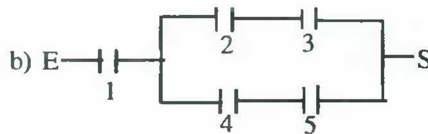
a) dos de los tres programas?.

b) al menos uno de los tres programas?. Rp. a) 0.29, b) 0.79

23. En una oficina hay dos computadoras A y B que trabajan de manera independiente. Si en un momento cualquiera la probabilidad de que la máquina B esté en mal estado es 1/4 y la probabilidad de que sólo la máquina A esté en mal estado es 3/10, ¿cuál es la probabilidad de que sólo la máquina B esté en malas condiciones?. Rp. 3/20.

24. En los circuitos de la figuras que siguen, la probabilidad de que cada llave se cierre (pase corriente) es p, $0 < p < 1$. Si todas las llaves se cierran o abren en forma independiente, calcular la probabilidad de que la corriente pase de E a S en a), y b).





Rp. a) $p(1+p-p^2)$, b) $2p^3-p^5$.

25. Un experimento se realiza tantas veces en forma independiente hasta obtener el primer éxito. Suponga que en cada intento la probabilidad de que se tenga éxito, es de 0.95 si se siguen correctamente las instrucciones; y es de 0.20 si no se siguen correctamente las instrucciones. Calcular la probabilidad de alcanzar el éxito en tres intentos a lo más

a) si se siguen correctamente las instrucciones cada vez,

b) si no se siguen correctamente las instrucciones cada vez.

Rp. a) 0.999875, b) 0.488.

26. Calcular la probabilidad de que un mensaje de n ($n \geq 1$) dígitos binarios, (0, y 1) sea incorrecto, si la probabilidad de recibir un dígito incorrecto es p y si los dígitos se reciben en forma independiente. Rp. $1-(1-p)^n$.

27. Suponga que un sistema funciona si al menos una de sus componentes funciona. Si las componentes trabajan independientemente y si la probabilidad que falle cada una es de 0.01, ¿cuántas componentes debería tener el sistema para que no falle con probabilidad de 0.9999?. Rp. 2.

28. Una persona está expuesta a 1 riesgo en 100 ocasiones independientes. Si la probabilidad de que ocurra un accidente es 1/100 cada vez, hallar la probabilidad de que un accidente ocurra en una o más ocasiones.

Rp. $1-(99/100)^{100} = 1-0.366 = 0.634$.

29. Un experimento aleatorio se repite sucesivamente 10 veces en forma independiente. En cada prueba la probabilidad de éxito es 1/4. Calcular la probabilidad de que ocurran 3 éxitos si el último intento debe ser un éxito.

Rp. $C_2^9(0.25)^3(0.75)^7$.

30. Respecto al partido de fútbol que protagonizarán los equipos A y B el próximo domingo, se piensa lo siguiente: De todas maneras se abrirá el marcador y cualquiera de los dos equipos tiene igual probabilidad de hacerlo. Si A anota el primer gol, la probabilidad de que el próximo también sea de A es 2/3 contra 1/3 de que sea de B; en cambio si B es el que anota, primero el gol, habrá un segundo gol que puede ser con igual probabilidad para cualquier bando. Si el marcador llega a ponerse dos a cero a favor de cualquier equipo la desmoralización de uno y la apatía del otro impedirán que haya más goles; en cambio si llega a ponerse 1-1, puede ocurrir tres cosas con iguales probabilidades: que A anote y gane 2-1, que B anote y gane 2-1 o que no haya más goles. Use un diagrama de árbol para calcular:

a) la probabilidad de que B gane.
b) la probabilidad de que B haya abierto el marcador dado que ganó el partido.

Rp. a) $P(B\text{ gane}) = \frac{1}{2}\times\frac{1}{3}\times\frac{1}{3}+\frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}\times\frac{1}{3} = \frac{14}{36}$, b) $(\frac{1}{4}+\frac{1}{12})/\frac{14}{36}$

31. Suponga que en cierta región del país la probabilidad de que un adulto mayor de 40 años tenga cáncer es 0.05. La probabilidad de que el diagnóstico sea correcto es 0.80, y de que sea errado es 0.20. Si se elige al azar a una de esas personas, calcular la probabilidad de que

a) se le diagnostique cáncer.
b) si se le diagnostica cáncer, tenga realmente tal enfermedad.

Rp. a) 0.23 b) 0.04/0.23 = 0.1739.

32. Ante una pregunta de opción múltiple de 5 alternativas donde sólo una es la respuesta correcta, un examinado, puede saber la respuesta o no saberla o tener dudas. Si no sabe, marca al azar. Si duda, reduce las alternativas a 3 de las cuales una es la correcta y luego, responde al azar. Si la probabilidad de que conozca la respuesta es 0.5, de que no conozca es 0.2 y de que dude es 0.3

a) Hallar la probabilidad de que acierte la pregunta.

b) Si acertó la pregunta, ¿que probabilidad hay de que no haya sabido la repuesta? Rp. a) 0.64 b) 0.04/0.64.

33. Sólo el 60% de la mercadería que recibe un comerciante del fabricante A es de calidad excepcional, mientras que el 90% de la mercadería que recibe del fabricante B es de calidad excepcional. Sin embargo la capacidad de fabricación del fabricante B es limitada, y por esta razón sólo el 30% de la mercadería le es permitido adquirir del fabricante B, el 70% la adquiere de A. Se inspecciona un embarque que acaba de llegar y se encuentra que es de calidad excepcional, ¿cuál es la probabilidad de que provenga del fabricante A?. Rp. 0.42/0.69.

34. En un proceso de producción el porcentaje de objetos no defectuosos fabricados es 70% con probabilidad 0.35, 90% con probabilidad 0.25, y 60% con probabilidad 0.4. Si se selecciona al azar uno de tales objetos y si resulta no defectuoso, calcular la probabilidad de que sea de calidad del 90% no defectuoso. Rp. 0.225/0.71.

35. El 100% de una población de electores se divide en tres estratos sociales excluyentes: baja, media y alta; de manera que la clase baja o media son el 90% del total, y la clase media o alta el 40% del total. De los primeros sondeos realizados para las próximas elecciones, se afirma que el porcentaje de electores que votarían por el candidato D puede ser:

30% de clase baja   50% de clase media   70% de clase alta

a) Si se elige un elector al azar y se encuentra que vota por D, ¿cuál es la probabilidad de que pertenezca a la clase alta?.

b) Si se escogen dos electores al azar, ¿qué probabilidad hay de que uno de ellos vote por D?. Rp. a) 0.175, b) 0.48.

36. Una máquina produce un tipo de objeto en distintos periodos. Si la máquina está bien ajustada en un periodo, el 80% de los objetos producidos pasan el control de calidad, de otro modo sólo pasan el 60%. Se ha determinado que el 90% de los periodos la máquina está bien ajustada. De los 25 objetos producidos en un solo periodo se escogen 3 al azar y a la vez para el control de calidad.

a) ¿Qué probabilidad hay que sólo 2 pasen el control de calidad?.

b) Si sólo 2 pasen el control de calidad ¿qué probabilidad se tiene que haya sido producido cuando la máquina trabaja en un periodo de buen ajuste?.

Rp. a) 960/2300, b) 855/960.

37. El departamento de créditos de una tienda comercial afirma que según sus experiencias pasadas la probabilidad de que el 20 % de los clientes que compran por más de \$50 es igual a 0.3 y que la probabilidad de que el 60% de los clientes compren por más de \$50 es igual a 0.7. Sin embargo al entrevistar a dos clientes al azar se encuentra que los dos compraron por más de \$50. En base a este resultado, ¿qué modificación acerca de las probabilidades 0.3 y 0.7 deberá hacer la tienda comercial?.

Rp.D:2 compran por más de \$50. $P(20\%)=0.3$, $P(60\%)=0.7$, $P(D)=0.3(0.2)^2+0.7(0.6)^2=$ $0.012+0.252=0.264$, $P(20\%/D)=0.045$, $P(60\%/D)=0.955$

38. A un candidato le han indicado que obtendría el 60% de los votos con probabilidad 0.2, el 45% de los votos con probabilidad de 0.3 y el 70% de los votos con probabilidad 0.5. Después de preguntarle a 4 personas se obtiene que 2 de ellas votarían por el candidato. A la luz de este resultado, ¿cuál es la probabilidad de que el candidato obtenga el 60% de los votos?.

Rp. $0.2\,C_2^4\,0.6^2 0.4^2+0.3\,C_2^4\,0.45^2 0.55^2+0.5\,C_2^4\,0.7^2 0.3^2=0.1095$, $0.06912/0.10955=0.63$.

39. Una agencia de publicidad observa que el 2% de los compradores potenciales de un producto ve su propaganda por periódico, el 20% ve dicha propaganda por televisión y el 1% ve los dos tipos de propaganda. Además de cada tres que ven la propaganda uno compra dicho producto y el 7.9% compran y no ven la propaganda

a) ¿Cuál es la probabilidad de que el comprador potencial compre dicho producto si no vio la propaganda?.
b) Si un comprador potencial compra el producto, ¿cuál es la probabilidad de que no haya visto la propaganda?

Rp. a) $0.079/0.79=0.1$, b) $0.21(1/3)+0.79(1/10)=0.149$, y $0.079/0.149$

40. Un gerente está a la espera de la llamada telefónica de 3 de sus clientes para realizar un negocio. La probabilidad de que lo llamen cualquiera de sus 3 clientes en forma independiente es 0.3. Además la probabilidad de realizar el negocio es de 0.20 si llama un cliente, es de 0.4 si llaman dos clientes, y es de 0.8 si llaman los 3 clientes. Si ninguno de los 3 le llama, no realiza el negocio.

a) calcular la probabilidad de que realice el negocio.
b) ¿cuántas llamadas de clientes es más probable que haya recibido el gerente sabiendo que realizó el negocio?.

Rp. Si A_i: "llaman i clientes ", $i=0,1,2,3$, entonces, $P(A_i)=C_i^3(0.3)^i(0.7)^{4-i}$.

Si B: "realiza negocio", $P(B/A_0)=0$, $P(B/A_1)=0.2$, $P(B/A_2)=0.4$, $P(B/A_3)=0.8$

a)$0+0.0882+0.0756+0.0216=0.1854$, b) 1 pues es $0.0882/.1854$.

Capítulo 6

VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

6.1 Variable aleatoria

Una variable estadística es una característica (cualitativa o cuantitativa) que se mide u observa en una población. Si la población es aleatoria y la característica es cuantitativa la variable estadística es denominada variable aleatoria. Las variables aleatorias se clasifican en discretas y continuas.

Definición. Se denomina **variable aleatoria,** a una variable estadística cuantitativa definida en un espacio muestral Ω.

Esto es, una variable aleatoria X es una función definida en Ω tal que a cada elemento $w \in \Omega$ le asocia el número real $x = X(w)$, (figura 6.1).

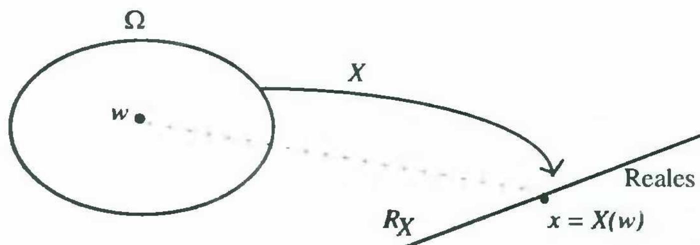


Fig. 6.1

El dominio de la variable aleatoria X es el espacio muestral Ω y el rango es un subconjunto de los números reales que denotaremos por R_X, siendo,

$$R_X = \{x \in \Re \,/\, x = X(w), w \in \Omega\}.$$

EJEMPLO 6.1.

Sea Ω el espacio muestral que se obtiene al lanzar al aire una moneda tres veces consecutivas, esto es,

$$\Omega = \{SSS, SSC, SCS, CSS, SCC, CSC, CCS, CCC\}.$$

Si X se define en Ω como "el número de caras obtenidas", entonces, X es una variable aleatoria cuyo rango es el conjunto: $R_X = \{0,1,2,3\}$. En efecto,

$X = 0$, corresponde al evento elemental $\{SSS\}$.

$X = 1$, corresponde a los eventos elementales $\{SSC\}$, $\{SCS\}$ y $\{CSS\}$.

$X = 2$, corresponde a los eventos elementales $\{SCC\}$, $\{CSC\}$ y $\{CCS\}$.

$X = 3$, corresponde al evento elemental $\{CCC\}$.

EJEMPLO 6.2.

Sea Ω, el espacio muestral que resulta de lanzar un dado sucesivamente veces hasta obtener el primer dos, esto es, .

$$\Omega = \{E, FE, FFE, FFFE, ...etc...\}$$

donde E(éxito)="sale dos", y F(fracaso)="No sale dos" , en cada prueba. La función Y definida en Ω, como "el número de lanzamientos realizados", es una variable aleatoria, cuyo rango es el conjunto: $R_Y = \{1,2,,etc..\}$. Esta variable ejecuta las siguientes asignaciones:

$$Y(E) = 1, \; Y(FE) = 2, \; Y(FFE) = 3, ... \text{ etc..}$$

En algunos casos el resultado w de un experimento aleatorio es ya la característica numérica que queremos anotar, en este caso se define la variable aleatoria X como la *función identidad*, esto es, $X(w) = w$.

EJEMPLO 6.3.

Sea Ω el espacio muestral que consiste del tiempo de duración de un artefacto electrodoméstico obtenido al azar de un lote, entonces, $\Omega = \{w \in \Re / w \geq 0\}$. Si la variable aleatoria X se define también como el tiempo de duración del artefacto, entonces X es la función identidad cuyo rango es el conjunto:

$$R_X = \{x \in \Re / x \geq 0\}.$$

NOTA. La variable aleatoria es pues, una *función* que atribuye a cada evento elemental *un número que no es aleatorio o imprevisible si no fijo y predeterminado.*

Lo que es aleatorio es el experimento sobre cuyo espacio muestral se define la variable aleatoria.

Esta función puede estar definida por una ley o regla o por un conjunto de números reales atribuidos mediante dicha ley o regla.

En la mayoría de las aplicaciones, se está más interesado en los valores que toma la variable, ignorando completamente el espacio muestral en el que se define.

El rango de la variable aleatoria **es un espacio** ***muestral inducido*** por la variable.

Clasificación.

Las variables aleatorias se clasifican en general en **discretas** y **continuas**.

Una **variable aleatoria discreta** es aquella cuyo rango es un conjunto finito o infinito numerable de valores (ejemplos 6.1 y 6.2).

Si la variable aleatoria X es discreta, **su rango se expresará en general** por:

$$R_X = \{x_1, x_2, \ldots, x_n \ldots\}.$$

Una **variable aleatoria continua** es aquella cuyo rango es un intervalo en $\Re$ o conjunto infinito no numerable de valores reales (ejemplo 6.3).

En general las variables aleatorias discretas representan datos que provienen del *conteo* del número de elementos, mientras que, las variables aleatorias continuas representan *mediciones*, por ejemplo, tiempo, peso, longitud, etc..

6.2. Variable aleatoria discreta: Función de probabilidad y función de distribución acumulada.

6.2.1. Probabilidad en el rango R_X.

Una variable aleatoria discreta asume cada uno de sus valores con cierta probabilidad que denotaremos por P_X (probabilidad inducida por X).

En efecto, si el rango de la variable aleatoria X es el conjunto finito de números, $R_X = \{1,2,\ldots,x_n\}$ y si $B = \{x_i\}$ es un evento en R_X, entonces,

$$P_X(\{x_i\}) = P(\{w \in \Omega \,/\, X(w) = x_i\}),$$

$$\text{o} \quad P_X(\{x_i\}) = P(A), \text{ donde, } A = \{w \in \Omega \,/\, X(w) \in B\}.$$

Con frecuencia, se utiliza la expresión $P[X = x_i]$ para denotar la probabilidad $P_X(\{x_i\})$. esto es,

$$P[X = x_i] = P(\{w \in \Omega \,/\, X(w) = x_i\}).$$

Por **ejemplo**, si la variable aleatoria X es el número de caras que resultan al tirar una moneda 3 veces, el rango de X es $R_X = \{0,1,2,3\}$, entonces,

$P[X = 0] = P(\{\text{SSS}\}) = 1/8.$

$P[X = 1] = P(\{\text{SSC o SCS o CSS}\}) = 3/8.$

$P[X = 2] = P(\{\text{SCC o CSC o CCS}\}) = 3/8.$

$P[X = 3] = P(\{\text{CCC}\}) = 1/8.$

En general, sea P una probabilidad definida en un espacio muestral Ω, y X una variable aleatoria definida en Ω cuyo rango es el conjunto de números R_X, la probabilidad P_X del evento B en R_X se define por:

$$P_X(B) = P(A)$$

siendo A en Ω un evento equivalente a B (figura 6.2), esto es,

$$A = \{w \in \Omega \,/\, X(w) \in B\}.$$

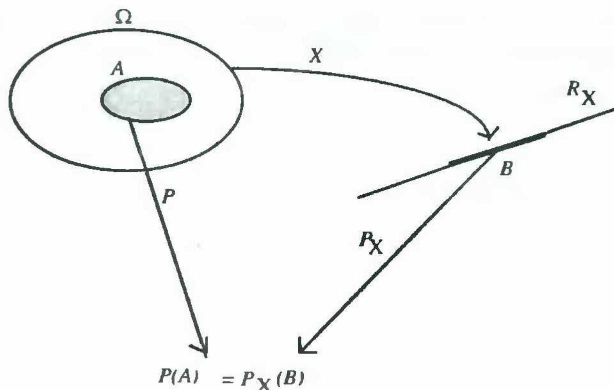


Fig. 6.2

NOTAS.

1. El conjunto de pares $(x_i, P[X = x_i])$ es la ***distribución de probabilidades*** de la variable aleatoria X.

Esta distribución **es similar a una distribución de frecuencias relativas**. por tanto, se pueden calcular, por ejemplo, medidas de tendencia central y de dispersión mediante un proceso similar al que se hizo con la distribución de frecuencias relativas.

Sólo habrá que cambiar las notaciones por tratarse ahora de una población.

2. Las probabilidades $p_i = P[X = x_i]$, $x_i \in R_X$ satisfacen las propiedades:

a) $p_i \geq 0$, para cada $x_i \in R_X$, b) $\sum_{x_i \in R_X} p_i = 1$.

3. Por extensión para todo número real $x \neq x_i$, siendo $x_i \in R_X$, se define:

$$P[X = x] = P(\emptyset) = 0$$

En consecuencia, queda definida la probabilidad $P[X = x]$ para todo número real x.

6.2.2 Función de probabilidad

Definición. Sea X una variable aleatoria discreta. Se denomina **función (ley o modelo o distribución) de probabilidad** de X a la función $f(x)$ definida por $f(x) = P[X = x]$ para todo x número real y que satisface las siguientes condiciones:

$$\text{i) } f(x) \geq 0 \quad \forall x \in \Re, \text{ y} \quad \text{ii) } \sum_{x_i \in R_X} f(x_i) = 1$$

La condición ii)

Es: $\sum_{i=1}^{n} f(x_i) = 1$, si $R_X = \{x_1, x_2, ..., x_n\}$ es finito.

Es: $\sum_{i=1}^{\infty} f(x_i) = 1$, si $R_X = \{x_1, x_2, ..., etc.\}$ es infinito,

NOTAS.

1. Si $A \subset R_X$, entonces, la probabilidad de A es el número:

$$P(A) = \sum_{x_i \in A} P[X = x_i] = \sum_{x_i \in A} f(x_i).$$

El lector debería verificar que $P(A)$ es una probabilidad, esto es, satisface los axiomas de probabilidad.

2. La función de probabilidad de una variable aleatoria X se puede expresar: por una ecuación : $f(x) = P[X = x] =$ expresión de x, o por el conjunto de pares $\{(x_i, p_i) / p_i = f(x_i),\ x_i \in R_X\}$ o por una tabla, como 6.1 (si R_X es finito).

Tabla 6-1: Distribución de probabilidad de v. a. discreta

Valores x_i de X	x_1	x_2	x_3	···	x_n
Probabilidad $p_i = P[X = x_i]$	p_1	p_2	p_3	···	p_n

La **gráfica** de una distribución de probabilidades discreta es la **gráfica de bastones** que consiste de segmentos verticales continuos o punteados de longitud proporcional a la probabilidad respectiva en cada valor x_i de la variable (figura. 6.3).

EJEMPLO 6.4.

Sea X la variable aleatoria definida como el número de caras que ocurren al lanzar una moneda 4 veces.

a) Determinar la distribución de probabilidades de X. Graficarla.

b) Calcular la probabilidad $P[0 < X \leq 2]$.

c) Determine la distribución de probabilidades de X si la moneda se lanza n veces ($n \geq 2$).

SOLUCION.

a) El rango de la variable aleatoria X, es el conjunto $R_X = \{0,1,2,3,4\}$. Suponiendo que los dieciséis sucesos elementales del espacio muestral son **equiprobables, la función de probabilidad,** es descrita por:

$f(0) = P[X = 0] = P(SSSS) = 1/16$

$f(1) = P[X = 1] = P(SSSC \text{ ó } SSCS \text{ ó } SCSS \text{ ó } CSSS) = 4/16$

$f(2) = P[X = 2] = P(SSCC \text{ ó } SCSC \text{ ó } SCCS \text{ ó } CSSC \text{ ó } CSCS \text{ ó } CCSS) = 6/16$

$f(3) = P[X = 3] = P(SCCC \text{ ó } CCSC \text{ ó } CSCC \text{ ó } CCCS) = 4/16$

$f(4) = P[X = 4] = P(CCCC) = 1/16$.

Observar que si $k \in R_X$, entonces, $X = k$, si y sólo si, en las 4 tiradas de la moneda aparecen k caras y $4 - k$ sellos. Esto ocurre de C_k^4 formas. Cada una de esas formas tiene probabilidad:

$$(1/2)^k (1/2)^{4-k} = (1/2)^4 = 1/16$$

siendo k = 0, 1, 2, 3, 4. Luego, la función de probabilidad del número de caras se puede describir como tabla 6.2 ó como la ecuación:

$$f(k) = \frac{C_k^4}{16}, \quad k = 0, 1, 2, 3, 4 .$$

Cuya gráfica es la figura 6.3.

b) $P[0 < X \leq 2] = \sum_{k=1}^{2} f(k) = f(1) + f(2) = \frac{4}{16} + \frac{6}{16} = \frac{10}{16}$.

c) $f(x) = C_x^n (1/2)^x (1/2)^{n-x} = \frac{C_x^n}{2^n}$, $x = 0, 1, 2, ..., n$

x_i	$f(x_i)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

Tabla 6.2

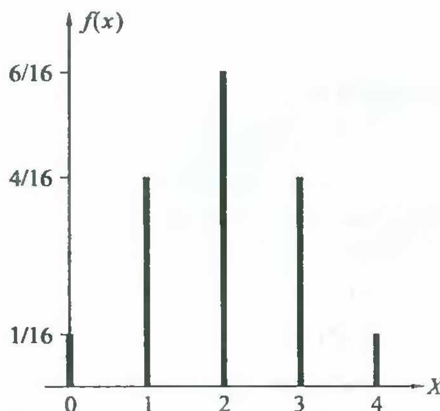


Fig. 6.3

NOTA. Si $p_i = f(x_i)$, la media de X se calcula por $\mu = \sum p_i x_i$, y la varianza por $\sigma^2 = \sum p_i x_i^2 - \mu^2$. En el ejemplo 6.4, el número de caras al lanzar una moneda 4 veces, tiene media $\mu = 2$ caras, y varianza $\sigma^2 = 1$ cara^2. (¡verificar!).

6.2.3 Función de distribución acumulada de variable aleatoria discreta

Definición. La **función de distribución acumulada** (f.d.a.) de probabilidades o simplemente **función de distribución,** $F(x)$, de la variable aleatoria discreta X, cuya función de probabilidad es $f(x)$, se define por:

$$F(x) = P[X \leq x] = \sum_{k \leq x} P[X = k] = \sum_{k \leq x} f(k), \quad \text{para} -\infty < x < \infty \cdot$$

EJEMPLO 6.5.

Sea X la variable aleatoria definida como el número de caras que resultan al lanzar una moneda 4 veces.

a) Hallar la función de distribución $F(x)$ de la variable aleatoria X y graficarla.

b) Usando $F(x)$, calcular $P[0 < X \leq 2]$.

SOLUCION.

a) La función de probabilidades $f(x)$ de la variable aleatoria X está descrita en el ejemplo 6.4, por:

$f(0) = 1/16$, $f(1) = 4/16$, $f(2) = 6/16$, $f(3) = 4/16$ y $f(4) = 1/16$.

Entonces, $F(0) = f(0) = 1/16$.

$$F(1) = f(0) + f(1) = 1/16 + 4/16 = 5/16$$

$$F(2) = f(0) + f(1) + f(2) = 1/16 + 4/16 + 6/16 = 11/16$$

$$F(3) = f(0) + f(1) + f(2) + f(3) = 1/16 + 4/16 + 6/16 + 4/16 = 15/16$$

$$F(4) = f(0)+f(1)+f(2)+f(3)+f(4) = 1/16+4/16+6/16+4/16+1/16 = 1$$

Por tanto,

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/16, & 0 \leq x < 1 \\ 5/16, & 1 \leq x < 2 \\ 11/16, & 2 \leq x < 3 \\ 15/16, & 3 \leq x < 4 \\ 1, & x \geq 4 \end{cases}$$

La gráfica de esta función de distribución está dada en la figura 6.4.

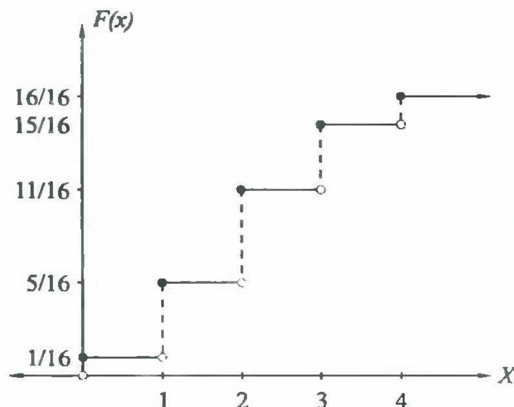


Fig. 6.4. Distribución de frecuencias acumuladas del número de caras

Observar que $F(x)$ da "saltos" en los valores de x=0,1,2,3,4. Por ejemplo, $F(2.99) = F(2) = 11/16$, $F(3) = F(3.99) = 15/16$, etc..

b) $P[0 < X \leq 2] = P[X \leq 2] - P[X \leq 0] = F(2) - F(0) = \frac{11}{16} - \frac{1}{16} = \frac{10}{16}$.

NOTA. En general si X tiene rango finito $R_X = \{x_1, x_2, ..., x_n\}$, y función de probabilidad $f(x_i)$, la función de distribución acumulada de X es:

$$F(x) = \begin{cases} 0, & -\infty < x < x_1 \\ \sum f(x_i), & x_i \leq x < x_{i+1}, \quad i = 1,2,...n-1 \\ 1, & x_n \leq x < +\infty \end{cases}$$

EJEMPLO 6.6.

Un lote de 10 artículos contiene 4 defectuosos. Si se obtiene una muestra al azar de cinco artículos, determine la distribución de probabilidades del número de artículos defectuosos en la muestra, si se escogen

a) Los cinco a la vez
b) Uno por uno con reposición

SOLUCION.

Sea X el número de artículos defectuosos en la muestra de cinco.

a) Si se escogen los 5 a la vez, el rango de X es el conjunto $R_X = \{0,1,2,3,4\}$. Si $k \in R_X$, el evento $X = k$ significa que de los 5 escogidos a la vez, k son defectuosos y $5-k$ son no defectuosos y esto ocurre de $C_k^4 C_{5-k}^6$. Por otro

lado, se escogen 5 artículos a la vez de 10, de C_5^{10} formas. Entonces, la función de probabilidad $f(x)$ de X es

$$f(x)=P[X=k]=\frac{C_k^4 C_{5-k}^6}{C_5^{10}},\ k=0,1,2,3,4.$$

El lector debería verificar que la suma de probabilidades es igual a 1.

También debería verificar que la misma distribución se obtiene si los cinco artículos se escogen **uno por uno sin reposición**.

b) Si se escogen los 5 uno por uno con reposición, el rango de X es el conjunto $R_X=\{0,1,2,3,4,5\}$. Si $k\in R_X$, $X=k$ significa que de los 5 artículos, k defectuosos ocurren independientemente con probabilidad $p=4/10$ cada uno, y $5-k$ no defectuosos ocurren independientemente con probabilidad $q=6/10$ cada uno. Estos eventos ocurren de C_k^5 formas. Entonces, la función de probabilidad $f(x)$ de X es

$$f(x)=P[X=k]=C_k^5 p^k q^{5-k},\ k=0,1,2,3,4,5.$$

donde $p=4/10$ y $q=6/10$.

El lector debe verificar que la suma de probabilidades es igual a 1.

EJEMPLO 6.7.

Una urna contiene 3 fichas de color rojo y una de color azul. Un experimento aleatorio consiste en extraer fichas al azar de la urna uno a uno sucesivamente.

a) Determinar la distribución de probabilidades del número de intentos que se realizan hasta que aparezca la primera ficha azul.
   a1) sin reposición, a2) con reposición.

b) Si dos personas A y B juegan sacando alternativamente una ficha con reposición de la urna y si gana el que obtiene la primera ficha azul, ¿cuál es la probabilidad de que A gane el juego si el juega primero?.

SOLUCION.

a) Sea X la variable aleatoria que denota el número de intentos hasta que ocurra la primera ficha azul.

   a1) Si la extracción es sin reposición, el rango de X es el conjunto $R_X=\{1,2,3,4\}$. Entonces,

$$f(1)=P[X=1]=\frac{1}{4}$$

$$f(2)=P[X=2]=\frac{3}{4}\times\frac{1}{3}=\frac{1}{4}$$

$$f(3)=P[X=3]=\frac{3}{4}\times\frac{2}{3}\times\frac{1}{2}=\frac{1}{4}$$

$$f(4)=P[X=4]=\frac{3}{4}\times\frac{2}{3}\times\frac{1}{2}\times\frac{1}{1}=\frac{1}{4}$$

Esta distribución de probabilidades iguales es conocida como **distribución uniforme**.

a2) Si la extracción es con reposición, el rango de X es el conjunto $R_X=\{1,2,3,\ldots,etc\}$. Si $k\in R_X$ entonces, $X=k$ sólo si ocurre la ficha azul en la k-ésima extracción con probabilidad $1/4$ y no azul en las anteriores extracciones independientes con probabilidad $3/4$. Luego. la probabilidad de $X=k$ está dada por:

$$f(k)=P[X=k]=\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^{k-1}, \qquad k=1,2,3,\ldots,etc.$$

Se verifica que:

$$\sum_{k=1}^{\infty} f(k)=1$$

En efecto, utilizando la suma:

$$\sum_{k=0}^{\infty} r^k=1+r+r^2+r^3+\ldots=\frac{1}{1-r},\quad \text{si}\ |r|<1$$

$$\sum_{k=1}^{\infty} f(k)=\frac{1}{4}\sum_{k=1}^{\infty}\left(\frac{3}{4}\right)^{k-1}=\frac{1}{4}\left(1+\frac{3}{4}+\left(\frac{3}{4}\right)^2+\left(\frac{3}{4}\right)^3+\ldots\right)=\frac{1/4}{1-(3/4)}=1\cdot$$

b) Si A sale primero. entonces. juega las veces impares. Luego,

$$P[\text{A gane el juego}]=P[X=1 \text{ ó } X=3 \text{ ó } X=5 \text{ ó } X=7, \text{etc..}]$$

$$=\left(\frac{1}{4}\right)+\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^4+\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^6+\ldots$$

$$=\left(\frac{1}{4}\right)\left[1+\frac{9}{16}+\left(\frac{9}{16}\right)^2+\left(\frac{9}{16}\right)^3+\ldots\right]$$

$$P[\text{A gane el juego}] = \left(\frac{1}{4}\right)\left[\frac{1}{1-(9/16)}\right] = \frac{4}{7}.$$

EJEMPLO 6.8.

En el ejemplo 6.7 determinar la distribución de probabilidades del número de intentos hasta que ocurra la segunda roja

a) Una por una sin reposición.
b) Una por una con reposición.

SOLUCION.

Sea X la variable aleatoria que denota el número de intentos hasta que ocurra la segunda ficha roja.

a) Si se escogen uno por uno sin reposición, el rango de X es el conjunto $R_X = \{2,3\}$. Entonces.

$$f(2) = P[X = 2] = P(RR) = \frac{3}{4}\times\frac{2}{3} = \frac{1}{2}$$

$$f(3) = P[X = 3] = P(RAR \text{ o } ARR) = \frac{3}{4}\times\frac{1}{3}\times\frac{2}{2} + \frac{1}{4}\times\frac{3}{3}\times\frac{2}{2} = \frac{1}{2}$$

b) Si se escogen uno por uno con reposición el rango de X es el conjunto $R_X = \{2,3,4,\ldots,etc\}$. Si $k \in R_X$, entonces, $X = k$ sólo si sale la segunda ficha roja en la k-ésima extracción con probabilidad $3/4$. En las $k-1$ extracciones restantes ocurre la otra roja con probabilidad $3/4$ y ocurren $k-2$ azules independientemente con probabilidad $1/4$. Luego, la probabilidad de $X = k$ está dada por:

$$f(k) = P[X = k] = (C_1^{k-1} q^{k-2} p)p = C_1^{k-1} q^{k-2} \text{p}^2 \qquad k = 2,3,..,etc.$$

donde $p = 3/4$, $q = 1/4$.

EJEMPLO 6.9.

En el ejemplo 6.7 si X es la variable aleatoria definida como el número de intentos que se realizan con reposición hasta que ocurra la primera ficha azul

a) Hallar la función de distribución acumulada $F(x)$ de X.
b) Utilizando $F(x)$ calcular $P[2 < X \leq 20]$.

SOLUCION.

a) La funcìon de distribución acumulada $F(x)$ de X está dada por la expresión:

$$F(x)=\begin{cases}0, & \text{si } -\infty < x < 1 \\ \sum_{t=1}^{x}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^{t-1}, & \text{si } k \le x < k+1,\end{cases} \quad , \ k=1,2,3,...$$

Observar que : $\sum_{t=1}^{x}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^{t-1} = 1-\left(\frac{3}{4}\right)^{x}$

b) $P[2 < X \le 20] = F(20) - F(2) = \left(\frac{3}{4}\right)^{2} - \left(\frac{3}{4}\right)^{20} = 0.559$

EJEMPLO 6.10.

Determinar el valor de la constante c para que la función $f(x)$ definida por:

$$f(x) = P[X = x] = \frac{c3^{x}}{x!}, \quad x = 0,1,2,...etc...$$

sea función de probabilidad de la variable aleatoria X, cuyos valores posibles son: 0, 1, 2,...,etc.. Hallar, además, la función de distribución acumulada de X y calcular $P[X \ge 3]$.

SOLUCION.

Cada probabilidad $f(x)$ debe ser mayor o igual que cero. Además, la suma de todas las probabilidades debe ser igual a uno.

Para encontrar la constante c utilizaremos la suma: $\sum_{k=0}^{\infty}\frac{z^{k}}{k!} = e^{z}$.

Luego, $\sum_{x=0}^{\infty}\frac{c3^{x}}{x!} = c\left(\sum_{x=0}^{\infty}\frac{3^{x}}{x!}\right) = ce^{3} = 1$, de donde resulta $c = e^{-3}$.

Entonces, $f(x) = \frac{e^{-3}3^{x}}{x!}, \quad x = 0,1,2,...etc...$

La función de distribución acumulada $F(x)$ de esta variable aleatoria es:

$$F(x)=\begin{cases}0, & \text{si } -\infty < x < 0 \\ \sum_{k=0}^{x}\frac{3^{k}e^{-3}}{k!}, & \text{si } t \le x < t+1,\end{cases} \quad , \ t=0,1,2,3,...$$

Por otra parte, $P[X \geq 3] = 1 - P[0 \leq X \leq 2] = 1 - [f(0) + f(1) + f(2)]$

$$= 1 - \left(e^{-3} + 3e^{-3} + 4.5e^{-3}\right) = 1 - 0.42319 = 0.57681 \cdot$$

EJEMPLO 6.11

Un inversionista puede participar en un negocio 4 veces y de manera independiente. Cada vez que participa gana o pierde $5000 con la misma probabilidad. La persona comienza con $10000 y dejará de participar en el negocio si pierde todo su dinero o gana $15000 (o sea si termina con $25000). Hallar la función de probabilidad de la utilidad del inversionista

SOLUCION

Sean: U la variable que representa la utilidad.

P = pierde con probabilidad = 1/2

G = pierde con probabilidad = 1/2

Entonces, los valores de U (montos finales menos capital) son:

−10,000, en los casos: PP o PGPP o GPPP con probabilidad=6/16

0, en los casos: PGPG o PGGP o GPPG o GPGP o GGPP con probabilidad =5/16

10,000, en los casos: GGPG o GPGG o PGGG con probabilidad=3/16

15,000, en los casos: GGG con probabilidad=2/16

6.3. Variable aleatoria continua: Función de densidad y función de distribución acumulada.

6.3.1 Función de densidad de probabilidad

Definición. Se dice que la función $f(x)$ es **función de densidad (ley** o **distribución) de probabilidad** de la variable aleatoria continua X si satisface las siguientes condiciones:

i) $f(x) \geq 0$ para todo $x \in \Re$ ($\Re$ es el conjunto de los números reales).

ii) $\int_{-\infty}^{+\infty} f(x)dx = 1$

iii) $P(A) = P[x \in A] = \int_A f(x)dx$, para cualquier intervalo $A \subset \Re$

NOTAS.

1. La condición i) expresa que la gráfica de $f(x)$ no tiene puntos por debajo del eje de las abscisas.

La condición ii), indica que el área total bajo la curva es igual a uno.

La condición iii) expresa: **probabilidad igual a área.** Esto es, si $A=[a,b]$, la probabilidad $P[a \leq X \leq b]$ es igual a la área de la región limitada por la curva, el eje X y las rectas $X = a$, $X = b$, es decir: (ver figura 6.5)

$$P[a \leq X \leq b] = \int_a^b f(x)dx$$

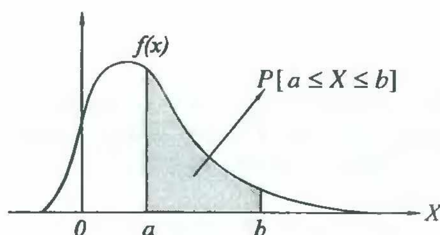


Fig. 6.5

2. No es difícil verificar que si el evento A es cualquier intervalo en la recta real, entonces, $P(A)$ satisface los axiomas de la probabilidad.

3. Cualquier función $f(x)$ que satisfaga sólo las condiciones i) y ii) no es probabilidad, *sólo es probabilidad cuando la función de densidad es integrada entre dos límites*. Por ejemplo, la función:

$$f(x) = 3x^2/8 \quad \text{para} \quad 0 \leq x \leq 2,$$

satisface las condiciones i) y ii), pero $f(x)$ no es probabilidad, ya que entre otros valores se tiene, $f(2) = 1.5$, $f(1.8) = 1.215$.

4. Si x_0 es cualquier valor específico de la variable aleatoria continua X, entonces:

$$P[X = x_0] = P[x_0 \leq X \leq x_0] = \int_{x_0}^{x_0} f(x)dx = 0$$

Como consecuencias se tiene:

a) $P(A) = 0$, no implica $A = \varnothing$.

b) $P[a \leq X \leq b] = P[a \leq X < b] = P[a < X \leq b] = P[a < X < b]$.

EJEMPLO 6.12.

Sea $f(x)$ una función definida en todos los números reales por

$$f(x)=\begin{cases} cx^2, & \text{si } x\in[0,2] \\ 0, & \text{si } x\notin[0,2] \end{cases}$$

a) Hallar el valor de la constante c para que, $f(x)$ sea una función de densidad para alguna variable aleatoria X.

b) Calcular $P[0 < X \leq 1]$.

SOLUCION.

a) El área de la figura 6.6 debe ser igual a 1, entonces,

$$1=\int_{-\infty}^{+\infty} f(x)dx=\int_0^2 cx^2dx=c\left[\frac{x^3}{3}\right]_0^2=c\frac{8}{3}$$

resultando, $c=3/8$. Luego,

$$f(x)=\begin{cases} 3x^2/8, & \text{si } x\in[0,2] \\ 0, & \text{si } x\notin[0,2] \end{cases}$$

La gráfica de esta función de densidad está dada en la figura 6.6.

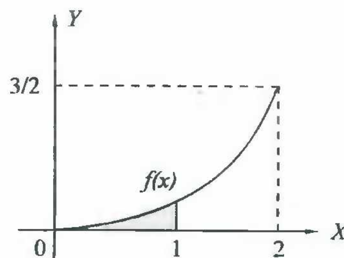


Fig. 6.6

b) $P[0 < X \leq 1]=\int_0^1 (\frac{3x^2}{8})dx=\left[\frac{x^3}{8}\right]_0^1=\frac{1}{8}$

6.3.2 Función de distribución acumulada de variable aleatoria continua

Definición. La **función de distribución acumulada (f.d.a),** $F(x)$ de una variable aleatoria continua X con función de densidad $f(x)$, se define (ver figura 6.7), por:

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)dt, \quad \text{para} -\infty < x < +\infty$$

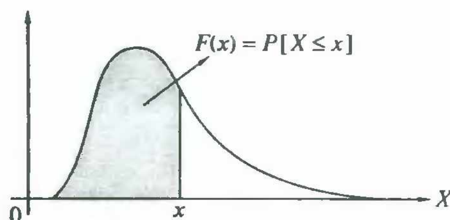
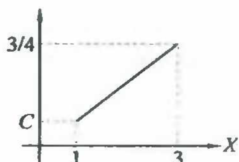


Fig. 6.7

EJEMPLO 6.13.

Sea $f(x)$ una función de densidad de probabilidad de una variable aleatoria X, cuya gráfica es la figura que sigue:



a) Determinar el valor de la constante c y luego la f.d.p. de X.
b) Hallar la función de distribución acumulada $F(x)$ de X.
c) Usando $F(x)$, calcular $P[5/4 < X \leq 5/2]$.

SOLUCION.

a) Si el área que encierra la figura con el eje X es igual a uno, entonces, $c = 1/4$.
La función de densidad se define a partir de la ecuación de la recta : $y - 1/4 = (3/4 - 1/4)/(3-1)(x-1)$, o $y = x/4$. Luego,

$$f(x) = \begin{cases} \dfrac{x}{4}, & \text{si } 1 \leq x \leq 3 \\ 0, & \text{en otro caso} \end{cases}$$

b) La función de distribución acumulada, se calcula por:

$F(x) = 0$, si $x < 1$, $\quad F(x) = 1$, si $x > 3$.

Si $1 \le x \le 3$, $F(x) = \int_1^x \frac{t}{4} dt = \frac{x^2}{8} - \frac{1}{8}$.

Entonces,

$$F(x) = \begin{cases} 0, & si\ x < 1 \\ \frac{x^2 - 1}{8}, & si\ 1 \le x \le 3 \\ 1, & si\ x > 3 \end{cases}$$

La gráfica de esta función de distribución acumulada es la figura 6-8.

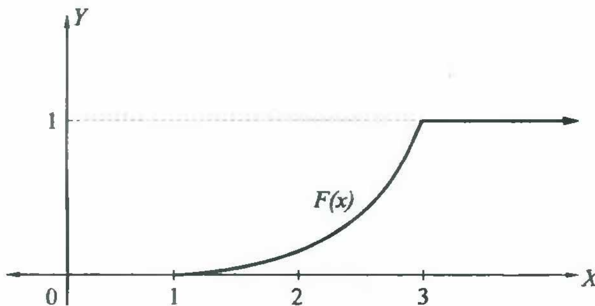


Fig. 6.8

c) $P\left[\frac{5}{4} < X \le \frac{5}{2}\right] = F\left(\frac{5}{2}\right) - F\left(\frac{5}{4}\right) = \frac{21}{32} - \frac{9}{128} = \frac{75}{128}$.

EJEMPLO 6.14.

La función de densidad de una variable aleatoria continua X, es descrita por la figura 6.9, es decir, $f(x) = 0$, si $x \notin [0,3]$, y,

$$f(x) = \begin{cases} cx, & \text{si } 0 \le x < 1 \\ c, & \text{si } 1 \le x < 2 \\ -cx + 3c, & \text{si } 2 \le x \le 3 \end{cases}$$

a) Determinar el valor de la constante c.
b) Hallar la función de distribución acumulada $F(x)$ de la variable aleatoria X y graficarla.

SOLUCION.

a) El área que encierra la figura 6.9 debe ser igual a 1, entonces, $c = 1/2$. También, si usamos integrales para el área se tiene:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^3 f(x)dx = c\int_0^1 xdx + c\int_1^2 dx + c\int_2^3 (-x+3)dx = 1$$

de donde resulta $c=1/2$. La función de densidad de X es entonces:

$$f(x) = \begin{cases} x/2, & \text{si } 0 < x \le 1 \\ 1/2, & \text{si } 1 < x \le 2 \\ (-x+3)/2, & \text{si } 2 < x \le 3 \\ 0, & \text{en otros casos} \end{cases}$$

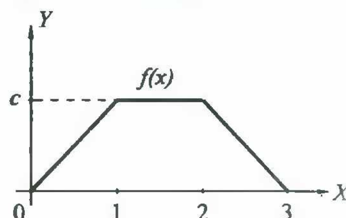


Fig. 6.9

b) Determinación de la función de distribución:

$$F(x) = \int_{-\infty}^{x} 0\,dt = 0, \text{ para } x \le 0$$

$$F(x) = F(0) + \int_0^x \frac{t}{2}dt = 0 + \left[\frac{t^2}{4}\right]_0^x = \frac{x^2}{4}, \quad \text{para} \quad 0 \le x \le 1,$$

$$F(x) = F(1) + \int_1^x \frac{1}{2}dt = \frac{(1)^2}{4} + \left[\frac{t}{2}\right]_1^x = \frac{x}{2} - \frac{1}{4}, \text{para } 1 \le x \le 2,$$

$$F(x) = F(2) + \int_2^x \left(\frac{-t}{2} + \frac{3}{2}\right)dt = \frac{3}{4} + \left[\frac{-t^2}{4} + \frac{3t}{2}\right]_2^x = -\frac{(x-3)^2}{4} + 1, \text{ para } 2 \le x \le 3,$$

$$F(x) = F(3) + \int_3^{+\infty} 0dx = 1 + 0 = 1, \text{ si } x > 3.$$

La función de distribución acumulada y su gráfica (fig. 6.10) son:

$$F(x) = \begin{cases} 0 & \text{si } -\infty < x \le 0 \\ x^2/4 & \text{si } 0 \le x \le 1 \\ (2x-1)/4 & \text{si } 1 \le x \le 2 \\ -\dfrac{(x-3)^2}{4} + 1 & \text{si } 2 \le x \le 3 \\ 1 & \text{si } 3 \le x < +\infty \end{cases}$$

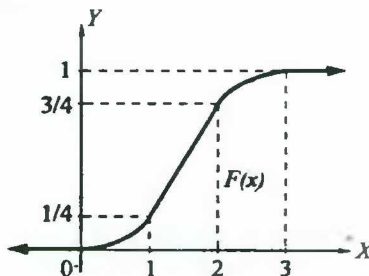


Fig. 6.10

EJEMPLO 6.15.

Si la función de densidad de una variable aleatoria continua X, es

$$f(x)=\begin{cases} ce^{-\beta x} & x \geq 0 \\ 0 & x<0 \end{cases}$$

a) Determinar el valor de la constante c.

b) Hallar la función de distribución acumulada $F(x)$ de la variable aleatoria X

SOLUCION

a) $1=\int_{-\infty}^{+\infty} f(x)dx=\int_{0}^{+\infty} ce^{-\beta x}dx = c\left[0+\frac{1}{\beta}\right]=c\frac{1}{\beta}$

resultando, $c=\beta$. Luego,

$$f(x)=\begin{cases} \beta e^{-\beta x} & x \geq 0 \\ 0 & x<0 \end{cases}$$

b) Si $x<0$, $F(x)=0$. Si $x \geq 0$, $F(x)=P[X \leq x]=\int_{0}^{x} \beta e^{-\beta t}dt = 1-e^{-\beta x}$

6.4. Propiedades de la función de distribución

1. La función de distribución acumulada $F(x)$ **es no decreciente**, esto es, para números reales x_1 y x_2 cualesquiera tales que $x_1 < x_2$, se tiene,

$$F(x_1) \leq F(x_2)$$

En efecto, sean los eventos $A=\{X \leq x_1\}, B=\{X \leq x_2\}$. Si $x_1<x_2$, entonces $A \subset B$. Luego, $P(A) \leq P(B)$ o $F(x_1) \leq F(x_2)$.

2. $\lim_{x \to -\infty} F(x)=0$ y $\lim_{x \to +\infty} F(x)=1$

A menudo, estos límites se escriben respectivamente: $F(-\infty)=0$ y $F(+\infty)=1$

En efecto, el evento $\{X \leq -\infty\}=\varnothing$ (es imposible), mientras que el evento $\{X \leq +\infty\}=\Omega$ (es seguro).

3. Para a, b números reales cualesquiera, tales que $a < b$, se tiene,

$$P[a < X \leq b] = F(b) - F(a).$$

En efecto, si $a < b$, entonces. $\{X \leq b\} = \{X \leq a\} + \{a < X \leq b\}$.
Por el axioma P3) de probabilidades, resulta,

$$P[X \leq b] = P[X \leq a] + P[a < X \leq b].$$

Por tanto, $P[a < X \leq b] = P[X \leq b] - P[X \leq a] = F(b) - F(a)$.

Observar que, si la variable aleatoria X es continua, entonces,

$$F(b) - F(a) = P[a < X \leq b] = P[a \leq X < b] = P[a < X < b] = P[a \leq X \leq b].$$

4. Dada la f.d.a. $F(x)$ de una variable discreta X, cuyos valores x_i, están ordenados en forma creciente, esto es,

$$x_1 < x_2 < x_3 < \ldots.$$

Entonces, la función de probabilidad $f(x)$ de X es determinada por:

$$f(x_i) = P[X = x_i] = F(x_i) - F(x_{i-1}), \quad i = 1,2,3,\ldots$$

donde, $F(x) = 0$ para todo $x < x_1$.

En efecto, si $x_1 < x_2 < x_3 < \ldots$, entonces,

$$P[X \leq x_i] = P[X \leq x_{i-1}] + P[X = x_i]$$

de donde resulta que para $i = 1,2,3,\ldots$

$$f(x_i) = P[X = x_i] = P[X \leq x_i] - P[X \leq x_{i-1}] = F(x_i) - F(x_{i-1})$$

Por ejemplo, si la f.d.a. de una variable aleatoria discreta X se define por:
$F(x) = 0$ si $x < 0$, $F(x) = 1/4$, si $0 \leq x < 1$, $F(x) = 3/4$, si $1 \leq x < 2$ y

$F(x) = 1$ si $x \geq 2$.

Entonces, la función de probabilidades $f(x)$ resulta:

$f(0) = F(0) - F(x_-) = 1/4 - 0 = 1/4$

$f(1) = F(1) - F(0) = 3/4 - 1/4 = 2/4$

$f(2) = F(2) - F(1) = 4/4 - 3/4 = 1/4$.

5. Dada la f.d.a. $F(x)$ de una variable aleatoria continua X, entonces su f.d.p $f(x)$ es igual a la derivada de la f.d.a. con respecto a x, donde ésta exista, esto es,

$$f(x)=\frac{d}{dx}F(x), \quad \forall x \quad \text{tal que} \quad \frac{d}{dx}F(x) \text{ exista}\cdot$$

En efecto, si $F(x)=\int_{-\infty}^{x} f(t)dt$, entonces, $f(x)=F'(x)$ para todo x donde exista la derivada $F'(x)$.

EJEMPLO 6.16.

Si la f.d.a. $F(x)$ de una variable aleatoria X se define por:

$$F(x)=\begin{cases} 1-ke^{-x/5}, & \text{si } x\geq 0 \\ 0, & \text{si } x<0 \end{cases}$$

a) Hallar la constante k y la función de densidad de probabilidad.
b) ¿Para qué valor de la constante c es $P[X\geq c]=0.01$?.

SOLUCION.

a) $F(+\infty)-F(0)=(1-k\times 0)-(1-k\times e^{0})=1$, entonces $k=1$. Luego,

$$f(x)=\frac{dF(x)}{dx}=\begin{cases} 1/5\,e^{-x/5}, & \text{si } x\geq 0 \\ 0, & \text{si } x<0 \end{cases}$$

b) $0.01=P[X\geq c]=1-P[X<c]=1-F(c)=1-(1-e^{-c/5})$.

De donde resulta , $c=23$.

EJEMPLO 6.17

Si la variable aleatoria continua X tiene función de densidad:

$$f(x)=\frac{1}{4}, \text{ si } 2\leq x\leq 6$$

y $f(x)=0$ en el resto. Determine la función de densidad de probabilidad de la variable aleatoria; $Y=3X-4$.

SOLUCION

$$G(y)=P[Y\leq y]=P[3X-4\leq y]=P[X\leq \frac{y+4}{3}]$$

$$G(y) = F(x) \text{, donde } x = \frac{y+4}{3}$$

$$g(y) = \frac{dG(y)}{dy} = \frac{dF(x)}{dx}\frac{dx}{dy} = f(x)\frac{1}{3}, \quad \text{donde } x = \frac{y+4}{3}.$$

Luego, $g(y) = \frac{1}{12}$, donde $2 \leq y \leq 14$.

EJEMPLO 6.18.

Un experimento consiste en seleccionar aleatoriamente un punto en el interior de una región triangular isósceles cuya base mide 6 cm. y cuyos lados iguales miden 5 cm. cada uno. Si X es la variable aleatoria que se define como la distancia del punto elegido a la base, hallar la función de densidad de X

SOLUCION.

Sea el triángulo ABC de la figura 6.11. La altura del triángulo es:

$$h = \sqrt{25-9} = 4$$

Como el punto se elige en el interior del triángulo, entonces el rango de X es el intervalo $R_X =]0, 4[$.

La f.d.a. $F(x)$ de X está dada por:

$$F(x) = P[X \leq x] = \begin{cases} 0, & \text{si} \quad x \leq 0 \\ \dfrac{\text{area del trapecio } AEFC}{\text{area del triángulo } ABC}, & \text{si } 0 < x < 4 \\ 1, & \text{si} \quad x \geq 4 \end{cases}$$

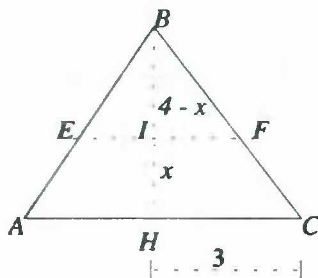


Fig. 6.11

De la figura 6.11 resulta,

Area del triángulo $ABC = 6 \times 4/2 = 12$.

Area del trapecio $AEFC = (x)\dfrac{(6+EF)}{2}$.

Por semejanza de los triángulos BHC y BIF se tiene:

$$\frac{IF}{HC} = \frac{BI}{BH} = \frac{\frac{EF}{2}}{3} = \frac{4-x}{4}$$

de donde se obtiene $EF = 6 - \dfrac{3x}{2}$.

Luego, la función de distribución de X es:

$$F(x) = \begin{cases} 0, & \text{si } x \le 0 \\ \dfrac{x}{2} - \dfrac{x^2}{16}, & \text{si } 0 < x < 4 \\ 1, & \text{si } x \ge 4 \end{cases}$$

Y la función de densidad de probabilidad es,

$$f(x) = \frac{dF(x)}{dx} = \begin{cases} \dfrac{1}{2} - \dfrac{x}{8}, & \text{si } 0 < x < 4 \\ 0, & \text{en otros casos} \end{cases}$$

EJERCICIOS

1. El número de hijos por familia en una determinada región es una variable aleatoria X cuya función de probabilidad es:

x	0	1	2	3	4
$P[X=x]$	1/16	4/16	k	4/16	1/16

a) Calcular el valor de la constante k.

b) Si una familia tiene al menos dos hijos, ¿cuál es la probabilidad de que tenga 3 hijos?.

Rp. a) $c=6/16$, b) $4/11$.

2. Una urna contiene 10 fichas de las cuales 4 son rojas y 6 son blancas. Se extraen 3 fichas al azar. Determine la distribución de probabilidades del número de fichas rojas. Si se escogen:

a) Los 3 a la vez

b) Una por una sin restitución

c) Una por una con restitución

Rp. a) X : 0, 1, 2, 3, $P[X=k]= C_k^4 C_{3-k}^6 / C_3^{10}$. b) X : 0, 1, 2, 3, $P[X=k]= C_k^4 C_{3-k}^6 / C_3^{10}$

c) X : 0, 1, 2, 3, $P[X=k]= C_k^3 p^k q^{3-k}$, $p=4/10$, $q=6/10$

3. Se venden 500 boletos de una rifa que consiste de un premio de \$200, 4 premios de \$50, y 10 premios de \$5. Si cada boleto cuesta 1\$, y si Ud. adquiere un boleto,

a) hallar la función de probabilidad de la utilidad.

b) qué probabilidad hay de ganar algún premio?.

Rp. a) valores: 199, 49, 4, −1, prob.: 1/500, 4/500, 10/500, 485/500, b) 0.03

4. Una caja contiene ocho focos de luz eléctrica, tres de los cuales son defectuosos. De la caja se selecciona al azar un foco y se la prueba, repitiéndose la operación hasta que aparezca un defectuoso. Sea X la variable aleatoria que se define como el número de extracciones necesarias hasta que aparezca el primer foco defectuoso.

a) Determinar la distribución de probabilidades de X, si las extracciones son sin reposición.

b) Hallar la función de distribución acumulada $F(x)$ de X si las extracciones son con reposición

Rp. a) valores: 1, 2, 3, 4, 5, 6, probab.: 21/56, 15/56, 10/56, 6/56, 3/56, 1/56

b) $F(x)=0$, si $x<0$, $F(x)=1-(5/8)^k$, $k\leq x<k+1$, $k=1,2,3$, etc

5. Una caja contiene 7 bolas de las cuales 4 son blancas. Se seleccionan estas bolas una por una y sin reemplazo. Sea X el número de selecciones necesarias hasta obtener todas las bolas blancas.
   a) Hallar la función de probabilidad de X.
   b) Calcular la probabilidad de que sean necesarios al menos seis selecciones.

   Rp. Valores de X: 4, 5, 6, 7, probabilidades: 1/35, 4/35, 10/35, 20/35, b) 30/35

6. Urna que contiene 3 fichas rojas y 5 azules. Un juego consiste en extraer una ficha sucesivamente son reposición. Si dos personas A y B juegan alternadamente extrayendo la ficha, hasta que ocurra una ficha azul. ¿cuál es la probabilidad de que A gane el juego si él sale primero?.

   Rp. X: # de intentos hasta obtener la primera azul. $P[X = k] = q^{k-1}p$, k=1,2,...etc. donde p=5/8, q=3/8. P(A gane)=P(X=1, X=3, X=5, etc.)=8/11

7. Se enumeran a 7 mujeres con los números 1, 2, 3, 4, 5, 6, 7 y del mismo modo se enumeran a 7 hombres con los números 1, 2, 3, 4, 5, 6, 7. Cada mujer en forma ordenada elige a un hombre al azar. Sea X la variable aleatoria que indica el número de mujer que elige al primer hombre con número impar.
   a) ¿Qué probabilidad existe de que sea la cuarta mujer la que elige al primer hombre impar?
   b) Halle la función de probabilidad de X.

   Rp. a) $P[X=4]=P[PPPI]=1/35$. b) X: 1, 2, 3, 4, probab : 20/35, 10/35, 4/35, 1/35

8. Una urna contiene 6 bolas numeradas de 1 a 6. Se extraen al azar dos bolas, una después de otra con reposición. Sea X el menor o igual de los dos números obtenidos.
   a) Encuentre la función de probabilidad de X.
   b) A partir de la función de distribución acumulada de X, calcular $P[2<X\leq 4]$.

   Rp. $f(x)=(13-2x)/36$, x=1,2,3,4,5,6, b)12/36

9. Se suelta un cuy en el centro de cuatro cajas colocadas en círculo, una de las cuales contiene un premio. El juego termina cuando el cuy entra en la caja que contiene el premio. Determinar la distribución de probabilidades del número de intentos hasta conseguir el premio si se supone que el cuy.
   a) tiene memoria, b) no tiene memoria

   Rp. a) valores: 1, 2, 3, 4, probabilidad : 1/4, 1/4, 1/4, 1/4, b) $P[X=k]=(3/4)^{k-1}(1/4)$, k=1, 2, 3, ...

10. Una urna contiene 3 bolas numeradas de 1 a 3. Un experimento aleatorio consiste en extraer las bolas una por una sin reposición. Si se considera un éxito cuando la bola k sale en la extracción k, $k = 1,2,3$,
    a) determine la función de probabilidad del número de éxitos..
    b) calcular la probabilidad de obtener al menos un éxito.

    Rp. a) valores: 0,1,3, probabilidad: 2/6, 3/6, 1/6, b) 4/6.

11. Un objeto producido puede contener, en forma independiente, a lo más tres tipos de defectos: A con probabilidad 0.04, B con probabilidad 0.08 y C con probabilidad 0.05. Si se selecciona al azar uno de tales objetos,

a) ¿qué probabilidad hay de que sea defectuoso?.

b) hallar la distribución de probabilidades del número de defectos del objeto.

Rp. a) 0.16096, b) valores: 0, 1, 2, y 3, prob: 0.83904, 0.15208, 0.00872 y 0.00016

12. Un vendedor puede visitar en un día uno o dos clientes con probabilidades 2/5 y 3/5 respectivamente. De cada visita en forma, independiente, puede resultar una venta por $500 con probabilidad 1/6 o ninguna venta con probabilidad 5/6. Si X son las ventas diarias, calcular la media y la varianza de X.

Rp. X :montos de ventas diarias. Valores: 0, 500, 1000, prob: 45/60, 14/60, 1/60

13. Del total de personas que se presentan para un puesto de trabajo el 60% son hombres y el resto mujeres. Aquellos que reúnen todos los requisitos para dicho puesto son el 40% de los hombres y el 50% de las mujeres. De tres personas que se presentan

a) Hallar la distribución de probabilidades del número de personas que cubren el puesto de trabajo.

b) ¿Cuál es la probabilidad de que al menos dos personas consigan el puesto de trabajo?.

Rp. a) $p=0.6\times0.4+0.4\times0.5=0.44$, $P[X=k]=C_k^3\, p^k q^{3-k}$, $k=0,1,2,3$. b) 0.41.

14. Una urna contiene n bolas numerados de 1 a n. Un juego consiste en extraer al azar tales bolas una por una con reposición hasta volver a encontrar la primera bola repetida. Hallar el modelo de probabilidad del número de bolas extraídas .

Rp. $P[X=k]=(k-1)(1-1/n)^{k-2}(1/n)^2$, $k=2, 3,$ etc ...

15. Un blanco circular de radio 1 se divide en 5 anillos circulares por medio de 5 discos concéntricos de radios: 1/5, 2/5, 3/5, 4/5, 1. Un jugador lanza un dardo al blanco, si el dardo alcanza el anillo circular comprendido entre los círculos de radios $k/5$ y $(k+1)/5$, $k = 0,1,2,3,4$; tiene k puntos y gana $5 - k$ dólares. Determinar la distribución de probabilidades

a) del puntaje del jugador.

b) de la utilidad del jugador.

Rp. a) Valores de X: 0,1,2,3,4, probab: 1/25, 3/25, 5/25, 7/25, 9/25, b) valores Util: 5,4,3,2,1.

16. Una tienda comercial tiene dos computadoras en stock el viernes en la mañana. La tienda puede recibir más computadoras sólo hasta el día lunes. Las probabilidades de que sean requeridas por los clientes 0, 1, 2, computadoras el día viernes son respectivamente: 0.5, 0.3, 0.2 y para el día sábado son respectivamente: 07, 0.2, 0.1. Si las demandas de los dos días son

independientes, determine la distribución de probabilidad del número de computadoras que quedan al finalizar el día sábado.

Rp. Valores: 0, 1, 2, probabilidades: 0.34, 0.31, 0.35.

Distribuciones continuas.

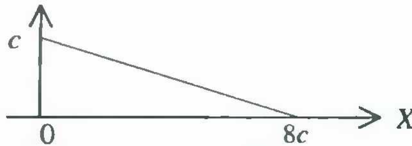
17. El tiempo de espera (en minutos) de un pasajero en un paradero de ómnibus en el intervalo [0, 5] es una variable aleatoria continua X cuya f.d.p. es :

$$f(x)=\begin{cases} c/5, & \text{si } 0\le x\le 5 \\ 0, & \text{en el resto} \end{cases}$$

a) Halle el valor c y la función de distribución acumulada $F(x)$ de X.
b) Calcule la probabilidad de que el pasajero espere al menos 2 minutos.
c) Calcule la probabilidad de que el pasajero espere exactamente 2 minutos.
d) ¿Cuánto es el tiempo máximo de espera para que tome el ómnibus con probabilidad 3/5?.

Rp. a) $c=1$, $F(x)=0$, si $x<0$; $F(x)=x/5$, si $0\le x\le 5$; $F(x)=1$, si $x\ge 5$, b) 3/5, c) 0, d) 3.

18. Suponga que el tiempo de vida de una componente electrónica, en miles de horas, es una variable aleatoria X cuya función de densidad tiene la gráfica siguiente,



a) Determinar c y la función de densidad de X.
b) Hallar la función de distribución acumulada de X.
c) Si el 95% de tales componentes duran a lo más k miles de hora, halle k.

Rp. a) $c=1/2$, $f(x)=(4-x)/8$ si $0\le x\le 4$, b) $F(x)=1-(x-4)^2/16$ si $0\le x\le 4$, c) 3.106 o 3106 horas.

19. La proporción de personas que contestan una encuesta por correo es una v.a. continua X con función de densidad de probabilidad:

$$f(x)=\begin{cases} c(2x+1), & \text{si } 0<x<1 \\ 0, & \text{en el resto} \end{cases}$$

a) Hallar el valor de la constante c y la función de distribución $F(x)$.
b) Si contestan al menos el 25%, ¿qué probabilidad hay de que contesten a lo más el 75%?.

Rp. a) $c=1/2$, $F(x)=(x^2+x)/2$ si $0<x<1$, b) calcular $P[X\le 0.75/X\ge 0.25]=16/27=0.59$.

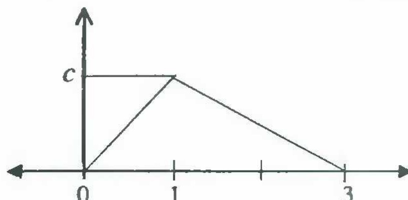
20. Suponga que el ingreso familiar mensual en miles de unidades monetarias (u.m.) en una ciudad, es una variable aleatoria X con función de densidad de probabilidad:

$$f(x)=\begin{cases} 4k, & \text{si } 0\le x\le 1 \\ k(5-x), & \text{si } 1\le x\le 5 \end{cases}$$

a) Determine el valor de la constante k.

b) Calcular el porcentaje de familias con ingresos mensuales de a lo más 2 mil u.m.. Rp. a) k=1/12, b) 62.5%.

21. La demanda semanal en miles de galones de gasolina en una estación de servicios, es una variable aleatoria X cuya función de densidad de probabilidad se aproxima a la función cuya gráfica es la figura que sigue:



a) Determinar la función de densidad de probabilidad de X.

b) Hallar la función de distribución acumulada de X. Graficar

c) ¿Qué cantidad de gasolina debe tener semanalmente la estación de servicio para satisfacer la demanda en el 62.5% de las semanas?

Rp. a) $c=2/3$, $f(x)=2x/3$, si $0\le x\le 1$, $f(x)=-(x-3)/3$ si $1\le x\le 3$, $f(x)=0$ en otros casos, c) 1.5.

22. Suponga que la demanda diaria de azúcar en un supermercado es una variable aleatoria X con función de densidad de probabilidad:

$$f(x)=\begin{cases} kx, & 0\le x\le 500 \\ k(1{,}000-x), & 500\le x\le 1000 \end{cases}$$

a) Determine el valor de la constante k.

b) Calcular la probabilidad de que en un día cualquiera el supermercado venda entre 250 y 750 kilogramos.

c) Calcular la cantidad Q de azúcar que debe ser dejada al público diariamente para que no falte azúcar en el 80% de los días.

Rp. a) $k=500^{-2}$, b) 0.75, c) 683.77.

23. El tiempo en horas que una familia mira televisión durante el día es una variable aleatoria X con función de densidad de probabilidad: :

$$f(x)=\begin{cases} kx, & si\ 0\le x<2 \\ 2k, & si\ 2\le x<6 \\ -k(x-8), & si\ 6\le x<8 \\ 0, & \text{en otros casos} \end{cases}$$

a) Hallar k y calcule la probabilidad de que la familia mire televisión entre 1 y 7 horas.

b) Determine la función de distribución acumulada $F(x)$. Graficar $F(x)$.

Rp. a) $k = 1/12$, $11/12$.

24. El tiempo de vida en meses de un producto perecible es una v.a. continua X con f.d.p.:

$$f(x)=\begin{cases} ce^{-\beta x} & si\ x\ge 0 \\ 0, & en\ otro\ caso \end{cases},\quad \beta>0$$

a) Determinar el valor de la constante c y la f.d.a. $F(x)$.

b) Si β=1/4, ¿qué probabilidad hay de que el producto dure al menos 4 meses?.

Rp. a) $c=\beta$, $F(x)=0$, si $x<0$; $F(x)=1-e^{-\beta x}$, si $x\ge 0$, b) 0.3679.

25. La cantidad de tiempo, en horas, que una computadora funciona antes de fallar, es una variable aleatoria X continua, cuya f.d.p. es:

$$f(x)=\begin{cases} ce^{-\frac{x}{200}}, & \text{si } x\ge 0 \\ 0, & \text{en otro caso} \end{cases}$$

Calcular la probabilidad de que la computadora funcione después de 100 horas.

Rp. $c = 1/200$, 0.6065

26. La duración, en decenas de horas, de cierta componente electrónica es una variable aleatoria continua X, cuya f.d.p es:

$$f(x)=\begin{cases} \dfrac{c}{x^2}, & \text{si } x\ge 100 \\ 0, & \text{en otro caso} \end{cases}$$

Si se prueban cuatro de estas componentes electrónicas, calcular la probabilidad de que dos de ellas duren más de 2,000 horas.(Suponga independencia en la duración de las componentes).

Rp. c=100 , p=P[X >200]=0.5, q=1−p. Prob.= $C_2^4 p^2 q^2 = 0.375$

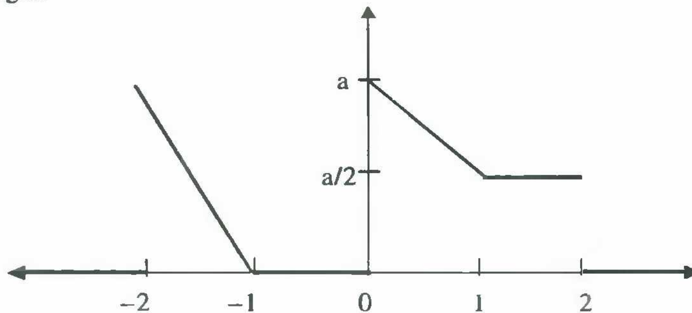
27. Sea X una variable aleatoria continua con función de densidad de probabilidad

$$f(x)=\begin{cases}\dfrac{1}{3}, & \text{si } 1 \le x \le 4\\ 0, & \text{en el resto}\end{cases}$$

determinar la función de densidad de probabilidad de la variable aleatoria $Y = 4X - 3$

Rp. G(y)=P[Y≤y]=P[4X−3≤y]= P[X≤(y+3)/4], g(y)=1/12, 1≤y≤13.

28. Si la *distribución de probabilidad mixta* de una v.a. X es descrita por la gráfica que sigue:



Hallar la f.d.p. $f(x)$ y la f.d.a. $F(x)$.

Rp. a=4/7. La f.d.p. es: $f(x)$=0, si x<−2, $f(x)$=−ax−a, si −2≤x<−1, $f(x)$ =0, si −1≤x<0, $f(x)$ = −(ax)/2 + a, si 0≤ x<1, $f(x)$=a/2, si 1≤ x<2, $f(x)$ = 0, si x>2.

6.5 Valor esperado o esperanza matemática.

Las distribución de probabilidad de una variable aleatoria se caracteriza básicamente a través de medidas de la tendencia central y de la dispersión. Estas medidas características de la distribución denominadas **parámetros** se describen por medio de la esperanza matemática.

6.5.1 Media de una variable aleatoria

La **media** de una variable aleatoria *X* o media de la distribución de probabilidad de *X* es un número real que se denota por μ_X o por μ. La media es denominada también **esperanza matemática** o **valor esperado** de *X*, y se denota también por *E*(*X*).

Definición 1. La media de una variable aleatoria discreta *X* con función de probabilidad $f(x)$ es la expresión:

$$\mu = E(X) = \sum_{x_i \in R_X} x_i f(x_i)$$

Si el rango de *X* es el conjunto finito $R_X = \{x_1, x_2, ... x_n\}$, entonces,

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

Si el rango de *X* es el conjunto infinito numerable $R_X = \{x_1, x_2, ...\}$, entonces,

$$E(X) = \sum_{i=1}^{\infty} x_i f(x_i)$$

En este caso, si la suma indicada no es igual a un número real, se dice que la esperanza de *X* no existe.

Definición 2. La media de una variable aleatoria continua *X* con función de densidad de probabilidad $f(x)$ es la expresión:

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx$$

siempre que el valor de la integral sea un número real, en caso contrario, se dice que la media de *X* no existe.

EJEMPLO 6.19

Calcular la media de la distribución de probabilidad de la variable aleatoria *X* que se define como el número de caras cuando se lanzan cuatro monedas.

SOLUCION.

La distribución de probabilidad de X se da en la tabla 6-3

Tabla 6-3

x_i	0	1	2	3	4
$f(x_i)$	1/16	4/16	6/16	4/16	1/16

La media de X es el número

$$\mu = E(X) = \sum x_i f(x_i) = 0\left(\frac{1}{16}\right) + 1\left(\frac{4}{16}\right) + 2\left(\frac{6}{16}\right) + 3\left(\frac{4}{16}\right) + 4\left(\frac{1}{16}\right) = 2$$

Esto significa que si una persona lanza 4 monedas, muchas veces, en promedio obtendrá 2 caras por lanzamiento.

NOTA. (Interpretación de la esperanza)

Refiriéndonos al ejemplo 6.19, supongamos que repetimos n veces el lanzamiento de las cuatro monedas y que se obtienen las frecuencias absolutas n_0, n_1, n_2, n_3 y n_4 de las veces que ocurren; 0, 1, 2, 3 y 4 caras respectivamente. Lo que resulta es una distribución de frecuencias cuyo promedio de caras por lanzamiento es igual a

$$\overline{X} = \frac{0\times n_0 + 1\times n_2 + 2\times n_3 + 3\times n_4 + 4\times n_4}{n} = 0\times\frac{n_0}{n} + 1\times\frac{n_1}{n} + 2\times\frac{n_2}{n} + 3\times\frac{n_3}{n} + 4\times\frac{n_4}{n}$$

En el cálculo de $E(X)$ se usan probabilidades o proporciones teóricas, mientras que en el cálculo de $\overline{X}$ se usan frecuencias relativas o proporciones empíricas obtenidas a partir de una muestra de tamaño n. A medida que n vaya creciendo es de esperar que las ***frecuencias relativas empíricas***:

$$n_0/n,\ n_1/n,\ n_2/n,\ n_3/n \ \text{ y } \ n_4/n$$

se vayan aproximando a las correspondientes probabilidades 1/16, 4/16, 6/16, 4/16, y 1/16. En consecuencia, **es de esperar que $\overline{X}$ se vaya aproximando a $E(X)$** a medida que n crece indefinidamente. Entonces, $E(X)$ es la media que se obtiene a "largo plazo" o a la "larga" , en otras palabras es la media que se espera obtener.

NOTA. (Mediana y moda de una variable aleatoria).

La **mediana** de una variable aleatoria X es el número Me tal que:

$$F(Me) = P[X \leq Me] = 0.5$$

Por **ejemplo**, acumulando las probabilidades ejemplo 6.19, resulta, $F(x)=5/16$, para $1 \le x < 2$, y $F(x)=11/16$, para $2 \le x < 3$.

Luego en $x=2$ puede ocurrir cualquier valor entre 5/16 y 11/16. En particular $F(2)=0.5$. Por lo tanto, la mediana=2.

Por otra parte, la **moda** de una variable aleatoria *X*, es el valor de la variable con mayor probabilidad (caso discreto), o donde alcanza su máximo (caso continuo). En la distribución del ejemplo 6.19, la moda=2.

EJEMPLO 6.20

Suponga que en un juego al azar consiste en lanzar un dado y que el jugador puede ganar $7, si obtiene al menos 5 puntos, o perder $2 en caso contrario.

a) ¿Cuánto espera ganar en el juego el jugador?.
b) ¿Cuánto debería ganar para que el juego sea justo?.

SOLUCION.

Sea la variable aleatoria *Y*, la cantidad de dinero que el jugador puede ganar. Entonces, los posibles valores de *Y* son: $7 si ocurre $A = \{5, 6\}$ con probabilidad 2/6, y −$2 si ocurre $A^C = \{1, 2, 3, 4\}$, con probabilidad 4/6.

a) El valor esperado de *Y* es, $E(Y) = -2 \times 4/6 + 7 \times 2/6 = 1$.

   En consecuencia, si el juego se repite indefinidamente, puede esperarse que el jugador gane en promedio $1.

b) El **juego es justo** si la esperanza es igual a cero, es decir si los esperados absolutos de ganar y de perder son iguales. Este juego será justo si el jugador gana sólo $4 dólares en el juego.

EJEMPLO 6.21

Suponga que el número de llamadas telefónicas que recibe una central en un período de tiempo, es una variable aleatoria *X*, con función de probabilidad:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0,1,2,\dots, etc.,\ \lambda = \text{constante}$$

Calcular la media del número de llamadas en ese período.

SOLUCION.

Utilizando la suma

$$1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots = \sum_{k=0}^{\infty} \frac{z^k}{k!} = e^z$$

se obtiene

$$E(X)=\sum_{x=0}^{\infty} xf(x)=\sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^{x}}{x!}=\lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda\cdot$$

EJEMPLO 6.22.

La vida útil de un objeto en miles de horas, es una variable aleatoria continua X cuya función de densidad de probabilidad es:

$$f(x)=\begin{cases}1-\dfrac{x}{2}, & \text{si } \ 0\leq x\leq 2\\ 0, & \text{en el resto}\end{cases}$$

Calcular la esperanza de vida del objeto.

SOLUCION.

$$E(X)=\int_{-\infty}^{+\infty} xf(x)dx=\int_{0}^{2} x\left(1-\frac{x}{2}\right)dx=\left[\frac{x^2}{2}-\frac{x^3}{6}\right]_0^2=\frac{2}{3}$$

En consecuencia, puede esperarse que la vida útil promedio del objeto sea de (2/3)1000 = 666.67 horas.

EJEMPLO 6.23.

Determine:

a) la moda, b) la mediana

de la distribución de la variable aleatoria X definida en el ejemplo 6.22.

SOLUCION

a) $f(0)=1$ es el valor máximo de la función de densidad (ver figura 6.12a), luego, la moda = 0.

b) La función de distribución acumulada de X es:

$$F(x)=0 \ \text{si} \ x<0, \quad F(x)=x-\frac{x^2}{4} \ \text{si} \ 0\leq x<2 \ \text{y} \ F(x)=1 \ \text{si} \ x\geq 2.$$

La mediana es el número Me que satisface

$$F(Me)=0.5 \ \text{ o } \ P[X\leq Me]=0.5$$

Luego, $$F(Me)=Me-\frac{(Me)^2}{4}=0.5, \ \text{con} \ 0\leq Me\leq 2$$

Resolviendo para Me se tiene: $Me=2\mp\sqrt{2}$.

Entonces, descartando el valor positivo, la mediana es $Me = 2-\sqrt{2} = 0.586$ (ver figura 6.12b).

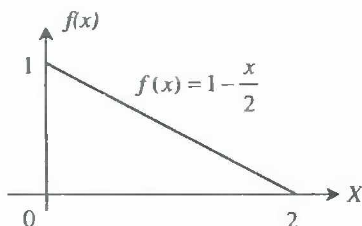


Figura 6.12 a)

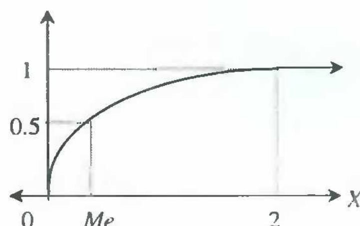


Figura 6.12 b)

6.5.2 Media de una función de una variable aleatoria

Sea X una variable aleatoria discreta con rango R_X y función de probabilidad $f(x) = P[X = x]$. Entonces, la función $Y = H(X)$, es una variable aleatoria con rango $R_Y = \{y / H(x) = y\}$, y con función de probabilidad $g(y)$ dada por:

$$g(y) = P[Y = y] = \sum_{\{x \in R_X \, / \, H(x) = y\}} P[X = x]$$

Por **ejemplo**, si X es una variable aleatoria cuya distribución de probabilidad está dada por la tabla 6.4, y si $Y = H(X) = 2X - 3$, entonces,

Tabla 6.4

x	0	1	2	3
$f(x)$	1/8	3/8	3/8	1/8

Y es una variable aleatoria, cuya distribución de probabilidades es dada por la tabla 6.5, donde $g(y) = P[Y = H(x)] = P[X = x] = f(x)$.

Tabla 6.5

$H(x) = 2x - 3$	-3	-1	1	3
$g(y) = f(x)$	1/8	3/8	3/8	1/8

El valor esperado de $H(x)$ es,

$$E(H(X)) = \frac{-3}{8} + \frac{-3}{8} + \frac{3}{8} + \frac{3}{8} = \sum H(x) f(x).$$

Definición. Si X es una variable aleatoria con distribución de probabilidad $f(x)$, **la media o valor esperado o esperanza matemática** de la variable aleatoria $H(X)$ está dada por la expresión:

$$E(H(X)) = \sum H(x_i) f(x_i), \quad \text{si } X \text{ es discreta,}$$

y por: $$E(H(X)) = \int_{-\infty}^{+\infty} H(x) f(x) dx, \quad \text{si } X \text{ es continua.}$$

Si la suma o la integral indicadas no son iguales a un número real se dice que la esperanza de $H(x)$ no existe.

EJEMPLO 6.24.

Suponga que un juego consiste en lanzar un dado y que si se obtiene al menos 5 puntos se gana $2, en caso contrario se pierde el número obtenido en dólares:

a) Defina la función utilidad en el juego.
b) Calcular la utilidad esperada en el juego.

SOLUCION.

Sea X la variable aleatoria definida como "el puntaje obtenido al lanzar el dado", entonces X toma los valores 1,2,3,4,5,6.

La distribución de probabilidades de X está dada por la tabla 6.7.

Tabla 6-7

k	1	2	3	4	5	6
$f(k)$	1/6	1/6	1/6	1/6	1/6	1/6

a) La función utilidad de este juego, es definida por:

$$U(x) = \begin{cases} 2 & \text{si } x = 5,6 \\ -x & \text{si } x = 1,2,3,4 \end{cases}$$

b) La esperanza de la función utilidad es igual a:

$$E(U(x)) = \sum_{x=1}^{6} U(x) f(x) = (-1)\times\frac{1}{6} + (-2)\times\frac{1}{6} + (-3)\times\frac{1}{6} + (-4)\times\frac{1}{6} + 2\times\frac{1}{6} + 2\times\frac{1}{6}.$$

$$E(U(x)) = -1$$

En consecuencia, si el juego se repite indefinidamente, puede esperarse que el jugador pierda en promedio **$1.**

EJEMPLO 6.25.

Una tienda de comestibles comercializa diariamente un producto que compra a \$8 y vende a \$10 cada unidad. Debido a que el producto es perecedero, las unidades que se queden sin vender al final del día, se desechan; perdiendo además del costo \$1 por unidad. El tendero ha establecido que la distribución de probabilidades de la demanda diaria del producto es la que se da en la tabla 6-8 .

Tabla 6-8

Demanda: d	0	10	20	30	40	50
Probabilidad	1/10	1/10	2/10	3/10	2/10	1/10

a) Si el tendero comercializara 30 unidades diariamente, ¿cuánto sería su utilidad esperada?.
b) ¿Cuántos unidades del producto debería comercializar diariamente a fin de maximizar su utilidad esperada?. Utilice sólo valores de la demanda

SOLUCION.

a) Si el tendero comercializa diariamente 30 unidades del producto, su *utilidad* para las demandas $d = 0, 10, 20, 30, 40, 50$, se da en la tabla 6-8b.

Tabla 6-8b

Demanda: d	0	10	20	30	40	50
Probabilidad	0.1	0.1	0.2	0.3	0.2	0.1
Utilidad	−270	−160	−50	60	60	60

Observar que el cálculo realizado se puede resumir en la **función utilidad**:

$$U = \begin{cases} 10\times 30 - 8\times 30, & \text{si } d \geq 30 \\ 10d - 8\times 30 - 1\times(30-d), & \text{si } d < 30 \end{cases} \qquad d = 0, 10, 20, 30, 40, 50$$

La utilidad esperada en el caso que el tendero comercialice 30 unidades es:

$$E(U) = (-270)\times 0.1 + (-160)\times 0.1 + (-50)\times 0.2 + 60\times 0.3 + 60\times 0.2 + 60\times 0.1 = -17$$

b) Si el tendero comercializa diariamente k unidades del producto, se tienen las siguientes esperanzas (verificar!): si k=0, $E(U)$=0, si k=10, $E(U)$=9, si k=20, $E(U)$=7, si k=30, $E(U)$=−17, si k=40, $E(U)$=−74, y si k=50, $E(U)$=−153.
El valor máximo de la utilidad esperada ocurre cuando k=10. Luego, el tendero debería comercializar diariamente 10 unidades del producto para maximizar su utilidad esperada.
El cálculo de las utilidades para $k = 0, 10, 20, 30, 40, 50$, se puede resumir en la **función utilidad** denotada por $U(k,d)$ o por U:

$$U = \begin{cases} 2k, & \text{si } d \geq k \\ 11d - 9k, & \text{si } d < k \end{cases} \qquad d = 0, 10, 20, 30, 40, 50$$

EJEMPLO 6.26

Suponga que la vida útil en años de cierto tipo de computadora es una variable aleatoria X con la función de densidad:

$$f(x)=\begin{cases} 2-2x & \text{si} \quad 0\le x\le 1 \\ 0 & \text{en el resto} \end{cases}$$

El distribuidor ofrece una garantía de 6 meses. Si la computadora falla en ese período se reemplaza por otra, a lo más una sola vez. Si cada computadora tiene el costo de fabricación de \$400 y el precio de venta de \$900, ¿cuánto es la utilidad esperada por computadora?

SOLUCION

Vida útil: X	$X\le 0.5$	$X>0.5$
Probabilidad	0.75	0.25
Utilidad	\$100	\$500

Utilidad esperada: $0.75\times 100+0.25\times 500=\$200$.

EJEMPLO 6.27

La demanda en miles metros de determinada tela que produce una compañía textil es una variable aleatoria X que tiene función de densidad siguiente:

$$f(x)=\frac{1}{10}, \quad \text{si} \quad 0\le x\le 10$$

Si por cada metro de tela vendida gana 4\$, pero por cada metro de tela no vendida en la temporada se pierde \$1, ¿cuál es la producción que maximiza la utilidad esperada de la compañía?.

SOLUCION

Sea k el valor de la producción tal que $0\le k\le 10$. Si x es el valor de la demanda, entonces, la función utilidad denotada por $U(x,k)$ o por U es :

$$U=\begin{cases} 4x-1\times(k-x) & \text{si} \quad 0\le x<k \\ 4k & \text{si} \quad k\le x\le 10 \end{cases}$$

La utilidad esperada de la compañía es entonces:

$$E(U)=\int_0^k (5x-k)f(x)dx+\int_k^{10} 4kf(x)dx=\int_0^k \frac{5x-k}{10}dx+\int_k^{10}\frac{4k}{10}dx$$

$E(U)=\dfrac{-5k^2}{20}+4k=\dfrac{-(k-8)^2+64}{4}$. Por gráfica, $E(U)$ es máximo en $k=8$.

También, por el método de la derivada $E(U)$ tiene valor máximo si $k=8$.

6.5.3 Varianza de una variable aleatoria

La varianza de una variable aleatoria X se denota por cualquiera de las formas: σ^2, σ_X^2, $Var(X)$, $V(X)$.

Definición. Sea X una variable aleatoria con distribución de probabilidad $f(x)$ y con media igual a μ. La **varianza** de X es la expresión:

$$\sigma_X^2 = E[(X-\mu)^2] = \sum (x_i-\mu)^2 f(x_i), \quad \text{si } X \text{ es discreta, y}$$

$$\sigma_X^2 = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx, \quad \text{si } X \text{ es continua.}$$

siempre que la suma y la integral sean números reales.

Definición. La desviación estándar de la variable aleatoria X es la raíz cuadrada positiva de su varianza. Esto es, $\sigma_X = \sqrt{\sigma_X^2}$

Una de las propiedades de la varianza que se usa en el cálculo es;

$$\text{var}(X) = E[(X-\mu)^2] = E(X^2) - \mu^2$$

EJEMPLO 6.28

Calcular la varianza y la desviación estándar de la distribución de probabilidad de la variable aleatoria X que se define como el número de caras al lanzar cuatro monedas.

SOLUCION.

La distribución de probabilidad de X es la tabla 6-9.

Tabla 6-9

X_i	0	1	2	3	4
$f(x_i)$	1/16	4/16	6/16	4/16	1/16

En el ejemplo 6.19, se ha calculado $E(X) = 2$. Además,

$$E(X^2) = \sum_{x=0}^{4} (x^2) f(x) = 0^2 \times \frac{1}{16} + 1^2 \times \frac{4}{16} + 2^2 \times \frac{6}{16} + 3^2 \times \frac{4}{16} + 4^2 \times \frac{1}{16} = \frac{80}{16} = 5$$

Por lo tanto, $\sigma_X^2 = E(X^2) - \mu^2 = 5 - (2)^2 = 1$

La desviación estándar de X es: $\sigma_X = \sqrt{Var(X)} = 1.$

EJEMPLO 6.29

Calcular la varianza y la desviación estándar de la variable aleatoria discreta X cuya función de probabilidad es:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \; x = 0,1,2,\ldots, etc., \; \lambda = \text{constante}$$

SOLUCION.

En el ejemplo 6.21 se ha calculado $\mu = E(X) = \lambda$. Además,

$$E(X^2) = \sum_{x=0}^{\infty} x^2 f(x) = \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} (k+1) \frac{\lambda^k}{k!}$$

$$= \lambda \left(e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} + e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) = \lambda \left(e^{-\lambda}(\lambda e^{\lambda}) + e^{-\lambda}(e^{\lambda}) \right) = \lambda^2 + \lambda.$$

Luego, la varianza de X es:

$$Var(X) = E\left(X^2\right) - \mu^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

La desviación estándar de X es

$$\sigma_X = \sqrt{Var(X)} = \sqrt{\lambda} = \lambda^{1/2}.$$

NOTA.

Otra forma de calcular esta varianza, es usando la siguiente expresión:

$$\sigma_X^2 = E[X(X-1)] + \mu - \mu^2$$

En efecto, $\mu = \lambda$, y

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2.$$

EJEMPLO 6.30

La vida útil de un objeto es una variable aleatoria X con función de densidad de probabilidad

$$f(x)=\begin{cases}\beta e^{-\beta x} & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x<0\end{cases}$$

calcular la varianza y la desviación estándar de X.

SOLUCION.

Integrando por partes y analizando la convergencia, se tiene:

$$\mu=\int_0^{+\infty} x\beta e^{-\beta x}dx=\underset{b\to+\infty}{Lim}\left[-xe^{-\beta x}\right]_0^b+\underset{b\to+\infty}{Lim}\left[-e^{-\beta x}\right]_0^b=0+\frac{1}{\beta}=\frac{1}{\beta}$$

El lector debería verificar (integrando dos veces por partes) que

$$E(X^2)=\int_0^{+\infty} x^2\beta e^{-\beta x}dx=\frac{2}{\beta^2}$$

Luego, $\sigma^2=E(X^2)-\mu^2=\dfrac{2}{\beta^2}-\dfrac{1}{\beta^2}=\dfrac{1}{\beta^2}$

La desviación estándar de X es: $\sigma=\dfrac{1}{\beta}$.

6.5.4 Variables aleatorias independientes

Definición. (función de probabilidad conjunta). Sean X e Y dos variables aleatorias discretas cuyos rangos son respectivamente

$$R_X=\{x_1,x_2,\ldots,x_n\} \qquad R_Y=\{y_1,y_2,\ldots,y_m\}$$

La probabilidad de que X e Y conjuntamente asuman los valores (x_i,y_i) se denomina **probabilidad conjunta** de X e Y.

La **función de probabilidad conjunta** de X e Y es la función $f(x,y)$ definida en el producto cartesiano $R_X \times R_Y$ por

$$f(x_i,y_i)=P[X=x_i \wedge Y=y_i], \text{ para todo } (x_i,y_i)\in R_X\times R_Y$$

Si se define $f(x, y) = 0$, para todo $(x, y) \in R_X \times R_Y$, entonces la función de probabilidad conjunta **satisface las propiedades**:

1) $f(x, y) \geq 0, \forall (x, y) \in \Re^2$

2) $\sum_x \sum_y f(x, y) = 1$

3) Para cualquier región A cn el plano XY, $P[(X, Y) \in A] = \sum\sum_A f(x, y)$

Por **ejemplo**, se seleccionan al azar dos fichas de una urna que contiene dos fichas rojas, tres blancas y dos azules. Si X es el número de fichas rojas e Y el número de fichas blancas, entonces, los posibles valores dc (X,Y) son: (0,0), (0,1), (1,0), (1,1), (0,2), (2,0) y la función de probabilidad $f(x,y)$ está dada por la tabla 6.10.

Tabla 6.10

$f(x,y)$	x=0	x=1	x=2	total fila: $h(y)$
y=0	1/21	2/21	3/21	6/21
y=1	6/21	6/21	0	12/21
y=2	3/21	0	0	3/21
Total columna: $g(x)$	10/21	8/21	3/21	21/21

Definición: Si $f(x,y)$ es la distribución conjunta de las variables X e Y. Las distribuciones marginales de X e Y están dadas respectivamente por:

$$g(x) = \sum_y f(x, y), \quad \text{y} \quad h(y) = \sum_x f(x, y).$$

En la tabla 6.10, $g(x)$ se ubica en el total columna, mientras que $h(y)$ se ubica en el total de fila.

Las funciones marginales para variables aleatorias continuas X e Y cuya función de densidad de probabilidad conjunta es $f(x,y)$, están dadas respectivamente por:

$$g(x) = \int_{R_Y} f(x, y)dy, \quad \text{y} \quad h(y) = \int_{R_X} f(x, y)dx$$

Definición. Se dice que dos variables aleatorias discretas X e Y son **estadísticamente independientes,** si los eventos: $[X=x_i]$ e $[Y=y_i]$ son independientes para cada par $(x_i, y_i) \in R_X \times R_Y$. Esto es, X e Y son dos variables aleatorias independientes, **si y** sólo si.

$$f(x_i, y_j) = P[X = x_i \wedge Y = y_j] = P[X = x_i]P[Y = y_j] = g(x_i)h(y_j)$$

para todo $(x_i, y_i) \in R_X \times R_Y$.

Definición. Se dice que dos variables aleatorias continuas X e Y son estadísticamente independientes, si y sólo si,

$$f(x, y) = g(x)h(y)$$

siendo $f(x,y)$ la función de densidad conjunta de las variables aleatorias X e Y, y $g(x)$, $h(y)$ las funciones marginales respectivas.

NOTA. En general, si $X_1, X_2,..., X_n$ son n variables aleatorias (discretas o continuas) independientes, entonces,

$$f(x_1, x_2,..., x_n) = f(x_1)f(x_2)...f(x_n).$$

NOTA. La función $H(X)$ de la variable aleatoria X, es una variable aleatoria. También la función $G(X,Y)$ de dos variables aleatorias X, Y, es una variable aleatoria. Por ejemplo, son variables aleatorias:

$$G(X,Y) = X + Y \quad G(X,Y) = X - Y\,,\ \ G(X,Y) = XY\,,\ \ G(X,Y) = X/Y\,.$$

6.5.5 Propiedades de la esperanza matemática

1. Si a y b son constantes reales, entonces, $E(aX + b) = aE(X) + b$.

2. Como consecuencias de la propiedad 1) se tienen:

2a) $E(b) = b$

2b) $E(X + b) = E(X) + b$

2c) $E(aX) = aE(X)$,

3. Si X e Y son variables aleatorias y a, b son constantes reales, entonces,

$$E(aX + bY) = aE(X) + bE(Y)\,.$$

En particular se tiene:

3a) $E(X + Y) = E(X) + E(Y)$, si $a = 1$ y $b = 1$.

3b) $E(X - Y) = E(X) - E(Y)$ si $a = 1$ y $b = -1$.

4. En general, si $X_1, X_2, ..., X_n$ son n variables aleatorias y $a_1, a_2, ..., a_n$ son n constantes reales, entonces,

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i).$$

5. Si X e Y son variables aleatorias independientes y a, b son constantes reales, entonces,

$$E(aXbY) = abE(X)E(Y).$$

6. En general, si $X_1, X_2, ..., X_n$ son n variables aleatorias independientes, entonces,

$$E(X_1 X_2 ... X_n) = E(X_1)E(X_2)...E(X_n).$$

6.5.6 Propiedades de la varianza

La varianza es un número real no negativo. Además se tiene:

1. $Var(X) = E(X^2) - \mu^2$

En efecto,

$$Var(X) = E[(X - \mu)]^2 = E[X^2 - 2\mu X + \mu^2]$$

Luego, $\sigma^2 = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2.$

2. Si a y b son constantes reales, entonces,

$$Var(aX + b) = a^2 Var(X).$$

3. Como consecuencias de la propiedad 2) se tienen:

3a) $Var(b) = 0$ si $a = 0$.

3b) $Var(X + b) = Var(X)$, si $a = 1$.

3c) $Var(aX) = a^2 Var(X)$

4. Si X e Y son dos variables aleatorias *independientes* y a, b son dos constantes reales, entonces,

$$Var(aX + bY) = a^2 V(X) + b^2 V(Y)$$

En efecto,

$$Var(aX+bY)=E[(aX+bY)^2]-[E(aX+bY)]^2$$

$$=E(a^2X^2+2abXY+b^2Y^2)-[E(aX)+E(bY)]^2$$

$$=a^2E(X^2)+2abE(XY)+b^2E(Y^2)$$
$$-a^2[E(X)]^2-2abE(X)E(Y)-b^2[E(Y)]^2$$

$$=a^2[E(X^2)-[E(X)]^2]+b^2[E(Y^2)-[E(Y)]^2]+0$$

$$=a^2Var(X)+b^2Var(Y).$$

En particular si X e Y son variables aleatorias independientes:

4a) $Var(X+Y)=Var(X)+Var(Y)$, si $a=1$ y $b=1$

4b) $Var(X-Y)=Var(X)+Var(Y)$, si $a=1$ y $b=-1$.

5.. En general, si $X_1, X_2, ..., X_n$ son n variables aleatorias independientes y $a_1, a_2, ..., a_n$ son n constantes reales, entonces,

$$Var\left[\sum_{i=1}^{n} a_i X_i\right]=\sum_{i=1}^{n} a_i^2 Var(X_i)$$

EJEMPLO 6.31

El número de defectos de un producto es una variable aleatoria X cuya distribución de probabilidad se da en la tabla 6-11,

Tabla 6-11

x	0	1	2	3	4	5
$f(x)$	1/10	2/10	3/10	2/10	1/10	1/10

Si el costo del producto está dado por la expresión.

$$C=4+2X-0.2X^2$$

determinar el costo esperado del producto.

SOLUCION.

$$\mu=0\times0.1+1\times0.2+2\times0.3+3\times0.2+4\times0.1+5\times0.1=2.3$$

$$E(X^2)=0^2\times0.1+1^2\times0.2+2^2\times0.3+3^2\times0.2+4^2\times0.1+5^2\times0.1=7.3$$

$$E(C) = 4 + 2E(X) - 0.2E(X^2) = 4 + 2\times 2.3 - 0.2\times 7.3 = 7.14$$

EJ EMPLO 6.32

Un dado se tira n veces. Calcular el valor esperado, y la varianza de la suma de los puntos que resulten.

SOLUCION.

Si X_i es el número de puntos que resulta en la tirada i, siendo i=1,2,3,...n. Entonces, la suma de estas n variables aleatorias independientes,

$$Y = \sum_{i=1}^{n} X_i$$

es la suma de los números de puntos que resultan de las n tiradas. Debemos calcular la esperanza y la varianza de la variable aleatoria Y.

La distribución de probabilidad de cada variable aleatoria X_i es dada en la tabla 6.12, de donde resulta: $E(X_i) = 7/2$.

Además, $E(X_i^2) = 91/6$,

Luego, $Var(X_i) = E(X_i^2) - \mu^2 = 91/6 - (7/2)^2 = 35/12$

Tabla 6.12

x	1	2	3	4	5	6
$f(x)$	1/6	1/6	1/6	1/6	1/6	1/6

Luego, el valor esperado de la suma de puntos es:

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{7}{2} = \frac{7n}{2}$$

y la varianza de la suma de puntos es:

$$V(Y) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} \frac{35}{12} = \frac{35n}{12}$$

EJEMPLO 6.33

La vida útil de una batería en años es una variable aleatoria X con la función de densidad de probabilidad.

$$f(x) = 0.2e^{-0.2x}, \qquad x \geq 0.$$

Si el costo del producto está dado por la expresión.

$$C = 80 + 2X + 0.2X^2$$

determinar el costo esperado de la batería.

SOLUCION

$E(X) = \dfrac{1}{0.2} = 5$. De $\sigma^2 = E(X^2) - \mu^2 = \dfrac{1}{(0.2)^2} = 25$, resulta: $E(X^2) = 50$.

Luego, $E(C) = 80 + 2E(X) + 0.2E(X^2) = 80 + 2\times5 + 0.2\times50 = 100$

EJERCICIO. (Desigualdad de Chebyshev).

Si X es cualquier variable aleatoria cuya distribución $f(x)$ tiene media μ y desviación estándar σ, pruebe que :

$$P[\mu - k\sigma \le X \le \mu + k\sigma] \ge 1 - (1/k^2), \quad k > 0$$

Prueba.

$$P[\mu - k\sigma \le X \le \mu + k\sigma] = \int_{\mu - k\sigma}^{\mu + k\sigma} f(x)dx = 1 - \left[\int_{-\infty}^{\mu - k\sigma} f(x)dx + \int_{\mu + k\sigma}^{+\infty} f(x)dx \right]$$

Por otra parte,

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{+\infty} (x-\mu)^2 f(x)dx$$

$$\sigma^2 \ge \int_{-\infty}^{\mu - k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu + k\sigma}^{+\infty} (x-\mu)^2 f(x)dx$$

$$\sigma^2 \ge k^2\sigma^2 [\int_{-\infty}^{\mu - k\sigma} f(x)dx + \int_{\mu + k\sigma}^{+\infty} f(x)dx], \text{ si } |x-\mu| \ge k\sigma \text{ ó } (x-\mu)^2 \ge k^2\sigma^2$$

De donde resulta: $\displaystyle\int_{-\infty}^{\mu - k\sigma} f(x)dx + \int_{\mu + k\sigma}^{+\infty} f(x)dx \le 1/k^2$.

NOTA. En particular si $k = 2$, $P[\mu - 2\sigma \le X \le \mu + 2\sigma] \ge 3/4$. Esto es, la probabilidad de que la variable aleatoria caiga dentro de dos desviaciones estándar de la media no puede ser menor de 3/4 o el 75%.

EJERCICIO. Si una variable aleatoria X tiene media =150 y varianza = 100, hallar una cota inferior para $P[100 \le X \le 200]$

SOLUCION. La cota mínima es 0.96. En efecto.

$$P[100 \le X \le 200] = P[|X - 150| \le 5\times10] \ge (1 - \frac{1}{5^2}) = 0.96$$

EJERCICIOS

1. Suponga que el número de cursos en que está matriculado un alumno es una variable aleatoria X con función de probabilidad;

x	1	2	3	4	5
$f(x)=P[X=x]$	1/10	2/10	3/10	3/10	1/10

a) Determinar la media y la varianza de la variable aleatoria X.
b) Si todos los cursos son de cuatro créditos, determinar la función de probabilidad, la media y la varianza del número de créditos.

Rp. a) 3.1, 1.29, b) 12.4, 20.64.

2. El número de defectos que puede tener una cerámica, es una variable aleatoria X con función de distribución acumulada $F(x)$ definida por la tabla:

X	$x<0$	$0 \le x < 1$	$1 \le x < 2$	$2 \le x < 3$	$x \ge 3$
$F(x) = P[X \le x]$	0.0	0.4	0.7	0.9	1.0

Hallar a) la media, b) la moda, c) la mediana y d) la varianza de X

Rp. a) 1, b) 0, c) 1, d) 1

3. Suponga que una lotería consiste de 1000 boletos que se venden cada uno a \$0.25, y de 5 premios respectivamente de \$25, \$20, \$10, \$5, \$1.
a) ¿Cuánto es la utilidad de la lotería si se venden los 1000 boletos?
b) Si Ud. juega esta lotería muchas veces, comprando un boleto cada vez, ¿cuánto espera ganar?.

Rp. a) \$189, a) –0.189.

4. La probabilidad de que un pequeño negocio sea destruido por incendio en el período de una año es de 0.005. Una compañía de seguros ofrece al dueño del negocio una póliza de seguro contra incendio en el termino de un año de \$10,000, cobrando una prima de \$250. ¿Cuánto es la utilidad esperada de la compañía de seguros?.

Rp. \$200.

5. Suponga que usted compra 3 cuyes escogiendo al azar uno por uno de una caja que contiene 4 cuyes hembras y 6 cuyes machos.
a) Hallar la media y la varianza del número de cuyes hembras escogidas.
b) Si paga S/.5 por cada cuy hembra y S/. 7 por cada cuy macho, hallar la media y la varianza de los pagos que usted realizaría.

Rp. a) 1.2, b) media=S/. 18.6, varianza=S/.2.24.

6 Se escriben los nombres de 4 hombres y 3 mujeres en pedacitos de papel y se les coloca en una caja. Si se selecciona al azar uno por uno y sin reemplazo los papelitos de esta caja y se define la variable aleatoria X como el número de selecciones hasta encontrar todos los nombres de los hombres.

a) Halle la función de probabilidad de X.

b) Calcule la media y la varianza de X

Rp. a) valores de X: 4, 5, 6, 7, probabilidades: 1/35, 4/35, 10/35, 20/35, b) μ=224/35, σ^2 =0.64

7. Un juego consiste en sacar una bola al azar sucesivamente y sin reposición de una urna que contiene cuatro bolas rojas y tres bolas blancas. Ud. pierde el juego si saca una bola roja.

a) ¿En cuántos intentos promedio pierde el juego?.

b) Por participar en el juego Ud. paga $6, y gana (en dólares) una cantidad igual a 10 menos 2 veces el número de intentos necesarias hasta obtener la bola roja. Si juega muchas veces, ¿cuánto espera ganar?.

Rp. a) 1.6, b) $0.8.

8. Una empresa saca al mercado cada sábado un nuevo producto cuya demanda semanal (en miles de unidades) es una variable aleatoria D con función de probabilidad:

$$f(d) = P[D = d] = c\frac{2^d}{d!},\ d = 1, 2, 3$$

a) Calcular el valor de c y la demanda esperada del producto.

b) El costo de producción semanal tiene un costo base fijo de $5,000 y un costo variable de $1 por unidad producida. La empresa recibe $8 por unidad vendida, lo que no se vende durante la semana se descarta sin pérdida alguna. Calcular la ganancia esperada de la empresa si cada semana produce 2,000 nuevas unidades.

c) ¿Cuántas unidades debería producir la empresa para maximizar su ganancia esperada?. Utilice sólo valores de la demanda.

Rp. a) c=3/16, (30/16)×1000 =1875, b) 6,000$, c) k=1 da 2000, k=2 da 6,000, k=3, da 70000

9. La demanda semanal de un artículo que produce una compañía varía grandemente de una semana a otra. De acuerdo con la experiencia, la demanda semanal y las respectivas probabilidades son las siguientes:

Demanda	2,000	3,000	4,000	5,000
Probabilidad	0.3	0.4	0.2	0.1

a) ¿Cuántas unidades de dicho artículo debe producir la compañía semanalmente, si decide programar la producción tomando como base:
a1) la demanda más probable?. a2) la demanda esperada?

b) Si cada unidad tiene un costo de S/. 5, se vende a S/.10 y si no se vende en la semana se debe desechar perdiendo S/.2.5 además del costo, ¿cuánto es la utilidad esperada si produce 3,000 unidades?.

c) ¿Cuántas unidades debe producir semanalmente la compañía a fin de maximizar su utilidad esperada?. Utilice sólo valores de la demanda

Rp. a1) 3000, a2) 3100, b) utilidad: 2500, 15000, 15000, 15000, Media=S/. 11,250 , c) 2,000 da S/.10,000, 3000 da S/.11,250, 4000 da S/.7500, 5,000 da 1250, la máxima para 3000

10. Usted ha arribado con dos maletas al control de equipajes de un aeropuerto junto con otros pasajeros. En el control se sigue la siguiente política: Se seleccionan al azar a 3 pasajeros para una revisión automatizada de sus maletas. Estos 3 pasajeros son ordenados al azar y sus maletas revisadas. La revisión de una maleta dura 3 minutos. Si usted es seleccionado y observa que los otros 2 pasajeros seleccionados tienen respectivamente 1 y 3 maletas, halle la media y la varianza del tiempo que usted se demorará en salir de este control

Rp. X tiempos : 6, 9, 15, 18. prob: 1/3, 1/6, 1/6, 1/3, $E(X)$=12, $V(X)$=27.

11. Un juego consiste en extraer al azar una ficha tras otra sin reposición de una urna que contiene 2 fichas rojas y 3 azules.

a) Hallar la ley de probabilidad del número de extracciones hasta obtener la segunda ficha azul.

b) Si por el número de extracciones X se gana $G=(40/X)-10$ hallar el valor esperado del juego, ¿sería usted un asiduo concurrente a este juego?

Rp. X:# de intentos hasta obtener 2da azul. a) Valores:2, 3, 4, prob: 36/120, 48/120, 36/120, b) G: 10, 10/3, 0, $E(G)$= 10×36/120+10/3×48/120+0=4.33, si.

12. Un comerciante recibe para su venta, un lote de cajas de 12 artículos cada una. El ha descubierto en anteriores entregas que hay un 25% de artículos defectuosos por caja. Su control de calidad por caja, consiste en escoger 3 artículos al azar uno a uno y sin reposición. Si los tres artículos escogidos no son defectuosos, no sigue probando el resto de la caja, pero si al menos uno de los 3 es defectuoso debe probar todos los artículos de la caja. Si se controlan muchas cajas, calcular el número esperado de artículos a probar por caja.

Rp. X: # artículos a probar, valores: 3, 12 , prob: 42/110, 68/110, $E(X)$=8.5636.

13. Un inspector A selecciona tres artículos al azar y a la vez de un lote de 50 artículos que contiene 5 defectuosos, anota el número de defectuosos hallados y repone los tres artículos al lote. Luego, el inspector B selecciona dos artículos uno por uno y sin reposición y anota el número de defectuosos hallados. Determine la media de la distribución de probabilidades del número total de defectuosos encontrados por los inspectores A y B

Rp. X: 0,1,2,3,4,5, prob: D=480200: 280962/D, 161865/D, 34023/D, 3213/D, 135/D 2/D, μ=0.5.

14. Una compañía se seguros propone asegurar cierta marca de automóvil del año en $20,000. La compañía estima que una pérdida total puede ocurrir con una probabilidad de 0.001, una pérdida del 50% con una probabilidad de 0.02, una

pérdida del 25% con una probabilidad de 0.2. Si se ignoran las demás pérdidas parciales, ¿qué prima debe cobrar la compañía de seguros si espera una ganancia de $200?. Rp. $1.420.

15. Un inversionista debe decidirse por uno de los dos proyectos: A o B. Si invierte en el proyecto A, puede ganar $41,000 si lo administra bien o perder $10,000 en caso contrario. Si invierte en el proyecto B, puede ganar $20,000 si lo administra bien o ganar $2,000 en caso contrario. Si las probabilidades son de 3 a 2 de que no lo administre bien,

a) ¿cuál de los proyectos debería tomar de manera que el beneficio esperado sea máximo?.

b) ¿cuánto debería ganar en el proyecto B si lo administra bien de manera que el beneficio esperado sea igual al del proyecto A?.

Rp. a) A, ya que espera ganar $10,400 contra $9,200 de B, b) 23,000.

Continuas

16. El costo C de un proyecto, en miles de dólares, está dado por: $C = 3X + 5X^2$, siendo X el tiempo empleado, una variable aleatoria que verifica:

$$E[(X - 1/2)^2] = 13/4, \text{ y } E[(X-1)^2] = 2$$

a) Determinar la media y la varianza de X.

b) Calcular el costo esperado del proyecto.

Rp. a) $E(X^2)=5$, $E(X)=2$, $Var(X)=1$, b) $31,000.

17. Una producto químico contiene una proporción X de cierto elemento por litro. El producto se vende a $6 litro si la proporción del elemento es menor de 50% y a $8 en caso contrario. Si el costo del producto es de $5 por litro y si X es una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} cx & 0 < x < 1 \\ 0 & \text{en otros casos} \end{cases}$$

hallar la función de probabilidad de la utilidad por litro y la utilidad promedio por litro.

Rp. c=2, valores: 1 y 3, probab: 0.25, 0.75, esperado $2.5.

18. Suponga que en cierta región el ingreso anual por persona en miles de dólares es una variable aleatoria X con función de densidad:

$$f(x) = \begin{cases} 4cx & \text{si } 0 \le x \le 1 \\ c(5-x) & \text{si } 1 \le x \le 5 \end{cases}$$

a) Hallar el valor de c.
b) Si se impone un impuesto de \$20 a los ingresos de menos de mil dólares, de \$50 a los ingresos de más de tres mil dólares y de \$40 a los otros ingresos, calcular la recaudación anual esperada por persona.
c) Si el impuesto es del 20% a los ingresos de menos de mil dólares, del 50% a los ingresos de más de tres mil dólares y del 40% a los otros ingresos, ¿cuánto es la recaudación anual esperada por persona?.

Rp. a) c=1/10, b) R: 20, 40, 50, prob.: 1/5, 3/5, 1/5, $E(R)$=\$380, c) $E(R)$=0.8466 o \$846.6

19. Sea X una variable aleatoria con función de densidad de probabilidad:

$$f(x)=\begin{cases}2x/9, & \text{si } 0\le x\le 3\\ 0, & \text{en otros casos}\end{cases}$$

a) Calcule la media, la mediana y la moda de X.
b) Halle la media y la varianza de la variable aleatoria. $Y = 2X - 3$.

Rp. a) media=2, mediana= $3/\sqrt{2}$, moda=3, b) $E(Y)$=1, $V(Y)$=2.

20. Suponga que el tiempo de atención (en minutos) a los clientes en la ventanilla de un banco, es una variable aleatoria X con función de densidad de probabilidad:

$$f(x)=\begin{cases}ce^{-x/3}, & \text{si } x\ge 0\\ 0, & \text{en otro caso}\end{cases}$$

a) Calcular la media μ y la desviación estándar σ de X.
b) Determinar la mediana de X
c) Calcular $P[\mu-2\sigma\le X\le\mu+2\sigma]$ directamente, y utilizando la desigualdad de Chebyshev.

Rp.c=1/3, a) 3, 3. b) 2.079. c) 0.95, ≥3/4.

21. Suponga que la vida útil (en horas) de cierta marca de "foco electrónico", es una variable aleatoria X cuya función de densidad de probabilidad es:

$$f(x)=\begin{cases}\dfrac{1}{8000}e^{-\frac{x}{8000}}, & \text{si } x\ge 0\\ 0, & \text{en otro caso}\end{cases}$$

a) Calcular la vida útil promedio por foco y la desviación estándar de X.
b) Si el costo C, en dólares, por cada foco está dado por:

$$C=20+\frac{X}{4{,}000,}+2\frac{X^2}{(8{,}000)^2}.$$

determine el costo promedio por foco.

Rp. a) 8,000, y 8,000. b) \$26

22. Un agente de bienes raíces cobra honorarios fijos de S/.200 más una comisión de 5% sobre el beneficio obtenido por el propietario. Si el beneficio del propietario es una variable aleatoria X con función de densidad de probabilidad

$$f(x)=\begin{cases}\dfrac{1}{2000} & 0\le x\le 2000\\ 0 & \text{en el resto}\end{cases}$$

a) ¿Cuánto espera obtener de utilidad el agente?
b) ¿Qué probabilidad hay de que obtenga utilidades superiores a S/.275?

Rp. a) $U=200+0.05X$, $E(U)=200+0.05E(X)=200+50=250$, b) $P[U>275]=P[X>1500]=0.25$

23. Un agente de bienes raíces gana S/.1000 más una fracción X de la ganancia que desea obtener el propietario. Si se considera a X como una variable aleatoria continua con función de densidad

$$f(x)=\begin{cases}Ax(1-x) & \text{si } 0\le x\le 1\\ 0 & \text{en el resto}\end{cases}$$

y un propietario quiere ganar 25000.
a) ¿Cuánto se espera que gane el agente?
b) ¿Cuál es la probabilidad de que el agente gane más de 11,000?.

Rp. $A=6$, $E(X)=0.5$, $G=1000+25000X$, a) $E(G)=1000+25000E(X)=13{,}500$,
b) $P[G>11000]=P[X>0.4]=0.648$.

24. La vida útil de una batería en años es una variable aleatoria X con la f.d.p.

$$f(x)=0.2e^{-0.2x},\qquad x\ge 0$$

El fabricante ofrece una garantía de un año. Si la batería falla en ese período se reemplaza por otra, a lo más una sola vez. Si el costo de fabricación es de \$20 por cada batería y se vende a \$50 cada una
a) ¿Cuál es la utilidad esperada del fabricante?
b) ¿Cuál debe ser el tiempo de garantía que el fabricante debe ofrecer para que sólo devuelva el 5% de las batería producidas?.

Rp. a) $30(e^{-0.2})+10(1-e^{-0.2})$, b) $0.05=P[X\le k]$, $k=0.2565$

25. La proporción de pureza X por galón de un compuesto que produce una empresa es una variable aleatoria con función de densidad:

$$f(x)=\begin{cases}3x^2 & 0\le x\le 1\\ 0 & \text{en otros casos}\end{cases}$$

Si la proporción de pureza del compuesto es menor o igual a un nivel k, el compuesto se destina a un fin A obteniéndose una utilidad de 50 dólares por galón. En caso contrario, el compuesto se destina a un fin B obteniéndose una utilidad de $60X$ dólares por galón. Determine el nivel k que le permita a la empresa maximizar su utilidad esperada por galón

Rp. Y: utilidad, $Y=50$ si $0\le x\le K$, $Y=60x$, si $k\le x\le 1$; $E(Y)=50k^3+45-45k^4$. $E(Y)$ es máximo si $K=150/180$.

26. La demanda en miles metros de determinada tela que tiene una compañía textil es una variable aleatoria X con función de densidad de probabilidad:

$$f(x)=\begin{cases}\dfrac{1}{5} & \text{si } \; 0\leq x\leq 5\\ 0 & \text{en el resto}\end{cases}$$

Por cada metro de tela vendida se gana 3$, pero por cada metro de tela no vendida en la temporada se pierde $1.Calcular la producción que maximiza la utilidad esperada de la compañía.

Rp. X: demanda, sea K la producción Y: utilidad, $Y=3K$ si $K\leq x\leq 5$, $Y=3x-1(K-x)$ si $0\leq x<K$. $E(Y)=(-2K^2/5)+3K$. $E(Y)$ es máximo si $K=15/4$.

27. La duración en horas de una tarjeta electrónica de una computadora es una variable aleatoria X con función de densidad: $f(x)=e^{-x}$, si $x\geq 0$. La computadora que usa esta componente produce un gasto de $\$a$ por hora para funcionar. Mientras la computadora está funcionando se obtiene una ganancia de $\$b$ por hora. Un operador de la computadora es contratado por H horas y se le paga $\$c$ por hora.

a) Defina la utilidad y determine la utilidad promedio.

b) ¿Para qué valor de H es máxima la utilidad promedio?.

Rp. a) U: utilidad $U=bx-ax-cH$ si $x<H$ y $U=bH-aH-cH$ si $x\geq H$

$E(U)=(b-a)[He^{-H}+1-e^{-H}(H+1)]-cH$, c) $H=-\ln(c/(b-a))$

28. Sea X una variable aleatoria con función de distribución acumulada

X	$x<0$	$0\leq x<1$	$1\leq x<3$	$x\geq 3$
$F(x)=P[X\leq x]$	0.0	$x^2/2$	$(x+1)/4$	1.0

Calcular la media y la varianza de X. Rp. $\mu=4/3$ y $\sigma^2=0.639$

29. Si X es una variable aleatoria con media 30 y varianza 9 usando el teorema de Chebyshev, hallar el valor de c tal que $P[|X-30|>c]<0.16$ Rp. $c=3\times 2.5$

30. Si X e Y son dos variables aleatorias independientes, probar que:

$$Var(X\pm Y)=Var(X)+Var(Y).$$

31. Si X es una variable aleatoria que tiene media μ y desviación estándar σ hallar la media y la varianza de la variable aleatoria estándar:

$$Z=(X-\mu)/\sigma$$

Rp. $E(Z)=0$, $Var(Z)=1$.

32. Sea X cualquier variable aleatoria. Si $E[(X-c)^2]$ existe, calcular el valor de la constante c que la haga mínima. Rp. $c=E(X)$.

Capítulo 7

ALGUNAS DISTRIBUCIONES IMPORTANTES

Introducción

El comportamiento de una variable aleatoria queda descrita por su distribución de probabilidad.

En muchas tareas estadísticas, se busca determinar una distribución de probabilidad o **modelo probabilístico** que satisfaga un conjunto de supuestos, para estudiar los resultados observados de un experimento aleatorio.

Se pueden definir muchas distribuciones de probabilidad tanto de variable aleatoria discreta como de variable aleatoria continua. Algunas, que tienen aplicaciones estadísticas importantes y que se estudiarán a continuación son:

Discretas:	**Continuas:**
Bernoullí	Uniforme
Binomial	Normal
Geométrica	Gamma
Pascal	Exponencial
Hipergeométrica	Chi-cuadrado
Poisson	t y F

7.1 ALGUNAS DISTRIBUCIONES IMPORTANTES DE VARIABLES ALEATORIAS DISCRETAS

7.1.1 Distribución de Bernoullí

Definición. Se denomina **prueba o ensayo de Bernoullí** a todo experimento aleatorio que consiste de sólo dos resultados posibles mutuamente excluyentes, generalmente llamados: **éxito** (E) y **fracaso** (F).

Por ejemplo, son ensayos de Bernoullí, lanzar una moneda al aire con los resultados: cara o sello. Elegir al azar un objeto fabricado, con los resultados: defectuoso o no defectuoso, etc..

El espacio muestral asociado al experimento aleatorio de Bernoullí se puede escribir como el conjunto: $\Omega = \{E, F\}$.

Definición. La variable aleatoria X definida en Ω de manera que atribuye a E el valor 1 y a F el valor 0, se denomina **variable aleatoria de Bernoullí.**

Definición. Si $p = P[X = 1]$ es la *probabilidad de éxito*, donde $0 < p < 1$, y $q = P[X = 0] = 1 - p$, es *la probabilidad de fracaso*, la **distribución de probabilidad de Bernoullí de parámetro** p, es descrita por la tabla:

X	0	1
$f(x)$	q	p

o por la ecuación:

$$f(x) = P[X = x] = p^x q^{1-x}, \quad x = 0,1.$$

La **función de distribución acumulada** de Bernoullí es:

$$F(x) = \begin{cases} 0, & \text{si } x < 0 \\ q, & \text{si } 0 \leq x < 1 \\ 1, & \text{si } x \geq 1 \end{cases}$$

TEOREMA 7.1.

Si X tiene distribución de Bernoullí de parámetro p, entonces,

a) $E(X) = p$, b) $Var(X) = pq$.

PRUEBA.

a) $\mu = E(X) = 0 \times (1 - p) + 1 \times p = p.$

b) $\sigma^2 = E(X^2) - \mu^2 = [(0)^2(1-p) + (1)^2(p)] - p^2 = p - p^2 = pq.$

NOTA.
la varianza de la distribución de Bernoullí es menor o igual que 1/4, en efecto,

$$\sigma^2 = p - p^2 = -\left(p - \frac{1}{2}\right)^2 + \frac{1}{4} \le \frac{1}{4}.$$

7.1.2 Distribución Binomial

Definición. Se denomina **experimento binomial** a un número fijo, n, de repeticiones independientes de un experimento aleatorio de Bernoullí, y por lo tanto, se caracteriza por que:

1. Las n pruebas son estadísticamente independientes.
2. Los resultados de cada prueba son dos mutuamente excluyentes, éxito (E) y fracaso (F).
3. La probabilidad p de éxito es invariante en cada una de las pruebas.

El espacio muestral del experimento binomial es el conjunto:

$$\Omega = \{(w_1, w_2, ..., w_n) / w_i = E \text{ ó } F\}.$$

Definición. Se denomina **variable binomial** a la variable aleatoria X definida en Ω como el número de éxitos que ocurren en las n pruebas de Bernoullí. Los posibles valores de X son: 0, 1, 2, 3, ..., n.

Si k es cualquier valor de la variable binomial, el evento $[X = k]$ consiste de todos los elementos de Ω que contengan k éxitos(E) y $n-k$ fracasos (F). La probabilidad de cada uno de estos eventos elementales es igual a $p^k q^{n-k}$. El número de estos eventos elementales es igual a:

$$\binom{n}{k} = C_k^n = \frac{n!}{k!(n-k)!}.$$

Por tanto, *la probabilidad de obtener k éxitos en n pruebas* Bernoullí es,

$$P[X = k] = \binom{n}{k} p^k q^{n-k}, \quad k = 0,1,2,3,...,n.$$

Para verificar que la suma de las probabilidades binomiales es igual a 1, se utiliza el binomio: $\sum_{k=0}^{n} C_k^n a^k (b)^{n-k} = (a+b)^n$ .En efecto,

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} = (p+(1-p))^n = (1)^n = 1.$$

Definición.- Se dice que **la variable aleatoria *X* definida como el número de éxitos que ocurren en las *n* pruebas de Bernoullí, tiene distribución binomial** con parámetros *n* y *p* y se escribe $X \sim B(n,p)$, si su función de probabilidad es:

$$f(k) = P[X = k] = \binom{n}{k}p^k q^{n-k}, \quad k = 0,1,2,3,...,n.$$

Observar que si $n = 1$, la función de probabilidad $B(1, p)$ es de Bernoullí.

La función de distribución acumulada de la variable aleatoria binomial es:

$$F(x) = P[X \le x] = \begin{cases} 0, & \text{si} \quad x < 0 \\ \sum_{k=0}^{x}\binom{n}{k}p^k q^{n-k}, & \text{si} \quad x = 0,1,2,...,n-1 \\ 1, & \text{si} \quad x \ge n \end{cases}$$

NOTA. Sean $X_1, X_2, ..., X_n$, *n* variables aleatorias independientes de Bernoullí con distribución $B(1, p)$ para cada $i = 1,2,3,...,n$. Entonces, la suma de estas *n* variables aleatorias independientes representa el **número de éxitos en *n* ensayos de Bernoullí**, esto es, la variable aleatoria:

$$X = \sum_{i=1}^{n} X_i$$

cuyos valores son: 0,1,2,..., *n*, tiene distribución Binomial $B(n, p)$.

TEOREMA 7.2

Si $X \sim B(n, p)$, entonces,

a) $\mu = E(X) = np$, b) $\sigma^2 = Var(X) = np(1-p)$.

PRUEBA.

La variable aleatoria *X*, número de éxitos en *n* ensayos de Bernoullí se puede escribir como una **suma de *n* variables aleatorias independientes** de Bernoullí X_i, esto es,

$$X = \sum_{i=1}^{n} X_i$$

siendo cada X_i de Bernoullí con: $E(X_i) = p$ y $Var(X_i) = p(1-p)$.
Luego,

a) $\mu = E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = np.$

b) $\sigma^2 = Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$

NOTA. Si $p = 1/2$, la distribución binomial $B(n, p)$ **es simétrica.** Además, si $p \to 1$, la distribución tiene **asimetría negativa** (cola a la izquierda), y si $p \to 0$, la distribución tiene **asimetría positiva** (cola a la derecha).

EJEMPLO 7.1 .

La probabilidad de que cierto tipo de objeto pase con éxito una determinada prueba es 5/6. Se prueban 10 de tales objetos. Si X es la variable aleatoria que se define como el número de objetos que no pasan la prueba:

a) Determine la función de probabilidades de X.
b) Calcule la media y la desviación estándar de X.
c) Determine la función de distribución acumulada $F(x)$ de X.
d) Usando $F(x)$, calcular $P[7 < X \leq 9]$.

SOLUCION.

a) Cada uno de los objetos puede no pasar la prueba (E) o puede pasar la prueba (F). La probabilidad de que el objeto no pase la prueba es 1/6 y de que pase la prueba es 5/6. La distribución de probabilidad X es entonces,

$$f(k) = P[X = k] = \binom{10}{k}\left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{10-k}, \quad k = 0,1,2,3,\ldots,10$$

b) La media de esta distribución es $\mu = np = 10(1/6) = 1.667$

Su desviación estándar es:

$$\sigma = \sqrt{np(1-p)} = \sqrt{10\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)} = 1.1785.$$

c) **Su función de distribución acumulada $F(x)$ es:**

$$F(x)=\begin{cases}0, & \text{si} \quad x<0 \\ \sum_{k=0}^{i}\binom{10}{k}\left(\frac{1}{6}\right)^{k}\left(\frac{5}{6}\right)^{10-k}, & \text{si} \quad i\leq x<i+1, i=0,1,2,3,\ldots,9 \\ 1, & \text{si} \quad x\geq 10\end{cases}$$

d) $P[7<X\leq 9]=F(9)-F(7)=\sum_{k=8}^{9}\binom{10}{k}\left(\frac{1}{6}\right)^{k}\left(\frac{5}{6}\right)^{10-k}=0.000.$

EJEMPLO 7.2.

En una tienda de alquiler de autos, cada vez que un cliente alquile un automóvil debe pagar como mínimo \$4. Si alquila un auto tipo A debe pagar \$15 más, y si alquila un auto tipo no A debe pagar \$5 más. Se sabe que la probabilidad de que un cliente alquile un auto tipo A es de 0.7. De cinco clientes que alquilan autos en esta tienda:

a) Determine la distribución de probabilidades de los clientes que alquilan automóviles tipo A.

b) Determine la utilidad y la utilidad esperada que producen a la tienda los 5 clientes que alquilan automóviles.

SOLUCION.

a) Sea X el número de clientes que alquilan automóviles tipo A. Entonces, los valores posibles de X son: 0, 1, 2, 3, 4, 5.

La probabilidad del evento, E: "un cliente alquila un automóvil tipo A" es $P(E)=0.7$ y $P(E^C)=0.3$. La distribución de probabilidad de X es:

$$f(k)=P[X=k]=\binom{5}{k}(0.7)^{k}(0.3)^{5-k}, \quad k=0,1,2,3,4,5.$$

b) La utilidad U que producen los cinco clientes es:

$$U=20+15X+(5-X)5, \quad x=0,1,2,3,4,5.$$

Dado que $E(X)=np=5\times 0.7=3.5$, la utilidad esperada es:

$$E(U)=45+10E(X)=45+10\times 3.5=80.$$

EJEMPLO 7.3.

Suponga que una producción de pollos bebé ha dado 10% de pollitos hembras.

a) Si la producción es llenada al azar en cajas de n pollos bebé cada una, hallar el valor de n de manera que la probabilidad de que no haya pollos bebé hembras en la caja sea igual a 0.08
b) Si cada caja contiene 24 pollos bebé, hallar la ley de probabilidad del número de pollos bebé hembras por caja, ¿cuál es el número esperado de pollitos hembras por caja?.
c) Si un criador de pollos recibe para su venta 20 cajas de 24 pollos bebé cada una, hallar el modelo de probabilidad del número de cajas que no contengan pollitos hembras, ¿cuál es el número esperado de cajas que no contengan pollos bebé hembras ?.

SOLUCION.

a) Sea X el número de pollos bebé hembras por caja de n pollos bebé. Los valores posibles de X son: 0, 1, 2,...,n.

La probabilidad de que un pollos bebé de la producción sea hembra es $p = 0.1$. Entonces, cada pollos bebé es colocado en la caja con probabilidad $p = 0.1$ de que sea hembra y con probabilidad $q = 1 - p = 0.9$ de que no sea hembra. Además las selecciones son independientes, ***esto se debe a que el número de* pollos bebé *producidos se considera muy grande***. Luego $X \sim B(n, 0.1)$.

Debemos hallar el valor de n tal que $P[X = 0] = 0.08$. Entonces, de

$$P[X = 0] = C_0^n (0.1)^0 (0.9)^n = 0.08 \text{, resulta } n = 23.976 \cong 24$$

b) Sea Y el número de pollitos hembras por caja de 24 pollitos. Los valores posibles de Y son: 0, 1, 2,...,24, Luego, $Y \sim B(24, 0.1)$, esto es:

$$P[Y = k] = C_k^{24} (0.1)^k (0.9)^{24-k}, \quad k = 0, 1, 2, 3, \ldots, 24.$$

El número esperado de pollitos hembras por caja es: $np = 24 \times 0.1 = 2.4$.

c) Sea W el número de cajas que no contengan pollitos hembras. Entonces, $W \sim B(20, p)$, donde p la probabilidad de que cada caja no contenga pollitos hembras en este caso es:

$$p = P[Y = 0] = C_0^{24} (0.1)^0 (0.9)^{24} = 0.0798$$

El número esperado de cajas que no contengan pollitos hembras es:

$$np = 20 \times 0.0798 = 1.596 \cong 2.$$

7.1.3 Distribución Geométrica

Definición. Se denomina **experimento geométrico** a las repeticiones independientes de un experimento aleatorio de Bernoullí hasta obtener el primer éxito. En cada ensayo de Bernoullí puede ocurrir un éxito (E) con probabilidad p o un fracaso (F) con probabilidad $q = 1 - p$, siendo $0 < p < 1$.

El espacio muestral del experimento geométrico es el conjunto:

$$\Omega = \{E, FE, FFE, \ldots, \}.$$

Definición. Se denomina **variable geométrica** a la variable aleatoria X definida como el número repeticiones independientes de un ensayo de Bernoullí hasta que resulte el primer éxito. Los posibles valores de X son: 1, 2, 3,..., etc..

Si k es uno de los valores de la variable, el evento, $[X = k]$ consiste del suceso elemental de Ω que contenga los primeros $k-1$ resultados fracasos y el último o k-ésimo resultado un éxito. *La probabilidad de que ocurra el primer éxito en la k-ésima prueba* es igual a $q^{k-1}p$, luego:

Definición. Se dice que **la variable aleatoria X que se define como el número de repeticiones independientes de un ensayo de Bernoullí hasta que ocurra el primer éxito, tiene distribución de probabilidad geométrica** con parámetro p y se escribe $X \sim G(p)$, si su función de probabilidad es:

$$f(x) = P[X = x] = q^{x-1}p, \qquad x = 1,2,\ldots,etc.$$

Para probar que la suma de las probabilidades geométricas es igual a 1, se utiliza la suma infinita:

$$\sum_{k=0}^{\infty} r^k = 1 + r + r^2 + r^3 + \ldots = \frac{1}{1-r}, \text{ si } |r| < 1.$$

En efecto, $$\sum_{k=1}^{\infty} q^{k-1}p = p(1 + q + q^2 + q^3 + \ldots) = p\left(\frac{1}{1-q}\right) = p\left(\frac{1}{p}\right) = 1$$

TEOREMA 7.3.

Si X es una variable aleatoria con distribución geométrica de parámetro p, entonces,

a) $\mu = \dfrac{1}{p}$, b) $\sigma^2 = \dfrac{q}{p^2}$

PRUEBA

a) Utilizando la identidad: $\displaystyle\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$ se obtiene:

$$E(X)=\sum_{k=1}^{\infty}kq^{k-1}p=p\sum_{k=1}^{\infty}kq^{k-1}=p\frac{1}{(1-q)^2}=p\frac{1}{p^2}=\frac{1}{p}.$$

b) Utilizando la identidad: $\sum_{k=1}^{\infty}k^2q^{k-1}=\frac{1+q}{(1-q)^3}$, se tiene:

$$E(X^2)=p\sum_{k=1}^{\infty}k^2q^{k-1}=p\left(\frac{1+q}{(1-q)^3}\right)=\frac{2-p}{p^2}$$

Luego, $Var(X)=E(X^2)-(E(X))^2=\frac{2-p}{p^2}-\frac{1}{p^2}=\frac{1-p}{p^2}=\frac{q}{p^2}$.

EJEMPLO 7.4

Pruebe que si X tiene distribución geométrica con parámetro p, entonces,

a) $P[X>a]=q^a$, $a\in Z^+$, $q=1-p$

b) $P[X>k+s/X>k]=P[X>s]$, $k,s\in Z^+$

SOLUCION

a) $P[X>a]=1-P[X\le a]=1-\sum_{k=1}^{a}q^{k-1}p=1-\frac{p}{q}\sum_{k=1}^{a}q^k=q^a$

b) $P[X>k+s/X>k]=\frac{P[X>k+s]}{P[X>k]}=\frac{q^{k+s}}{q^k}=q^s=P[X>s]$

EJEMPLO 7.5

Un vendedor a domicilio hace llamadas telefónicas a clientes potenciales. La probabilidad de vender en cada llamada es 0.02.

a) Calcule la probabilidad de que la sexta llamada sea su primera venta.
b) Calcule el esperado del número de llamadas hasta obtener su primera venta
c) ¿Qué probabilidad hay de que su primera venta ocurra después de más de 5 llamadas si ya se hizo 3 llamadas sin éxito?
d) El costo de cada llamada es S/.1.5. Si una llamada no es una venta la siguiente cuesta S/.0.5 más. Halle el costo esperado del número de llamadas que hace hasta que obtenga su primera venta.

SOLUCION.

Sea X el número de llamadas hasta conseguir una venta. Sus posibles valores son: 1,2,3,...,etc.. El modelo de probabilidad de X es geométrica de parámetro $p=0.02$, esto es:

$$P[X=k]=(0.02)(0.98)^{k-1},\quad k=1,2,3,...,$$

a) Luego, la probabilidad de que la sexta llamada sea su primera venta es:

$$P[X=6]=(0.02)(0.98)^5=0.018$$

b) $E(X) = 1/(0.02) = 50$. A la larga en la llamada #50 obtiene su primera venta

c) El evento "ya hizo 3 llamadas sin éxito" es equivalente al evento "requiere hacer más de 3 llamadas hasta que obtenga un éxito". Entonces,

$$P[X>5/X>3]=\frac{P[X>3\wedge X>5]}{P[X>3]}=\frac{P[X>5]}{P[X>3]}=\frac{(0.98)^5}{(0.98)^3}=(0.98)^2$$

d) Costo $=1.5+2\times(X-1)$, Costo esperado $=1.5+2\times((E(X)-1)=$ S/. 99.5

7.1.4. Distribución de Pascal o binomial negativa

Definición. Se denomina **experimento binomial negativo o de Pascal** a las repeticiones independientes de un experimento aleatorio de Bernoullí hasta obtener el éxito número *r*. En cada ensayo de Bernoullí puede ocurrir un éxito con probabilidad p o un fracaso con probabilidad $q=1-p$.

A la variable aleatoria *X* que se define como el número de intentos hasta que ocurra el éxito número *r* se le denomina **variable aleatoria binomial negativa o de Pascal**. Su rango es el conjunto: $R_X=\{r,r+1,r+2,...\}$.

Si $k\in R_X$, el evento $[X=k]$ ocurre, si resulta éxito en la *k*-ésima prueba y en las restantes $k-1$ pruebas resultan $r-1$ éxitos y $(k-1)-(r-1)=k-r$ fracasos. Luego, *la probabilidad de tener el r-ésimo éxito en la k-ésima prueba* es igual a:

$$C_{r-1}^{k-1}p^{r-1}q^{k-r}p=C_{r-1}^{k-1}p^r q^{k-r}$$

Definición. Si se repite en forma independiente un ensayo de Bernoullí, donde cada resultado puede ser un éxito con probabilidad *p* o un fracaso con probabilidad $q=1-p$, se dice que **la variable aleatoria *X* que se define como el número de intentos hasta que ocurran *r* éxitos**, tiene distribución binomial negativa o de Pascal con parámetros *r* y *p* y se escribe $X\sim P(r,p)$, si su función de probabilidad es:

$$f(k)=P[Y=k]=C_{r-1}^{k-1}p^r q^{k-r},\qquad k=r,r+1,r+2,...,etc.$$

Para verificar que la suma de las probabilidades de Pascal es igual a uno, se usa la siguiente serie binomial negativa:

$$(1-q)^{-r}=\sum_{i=0}^{\infty}C_i^{-r}(-q)^i=\sum_{k=r}^{\infty}C_{r-1}^{k-1}q^{k-r}.$$

NOTA. Sean $X_1, X_2, ..., X_r$, r variables aleatorias independientes cada una con distribución geométrica $G(p)$, tales que:

X_1 es el número de repeticiones necesarias hasta la ocurrencia del primer éxito (E).

X_2 es el número de repeticiones entre la primera ocurrencia de E hasta la segunda ocurrencia de E.

... Etc.

X_r es el número de repeticiones necesarias entre la $r-1$ ocurrencia de E hasta la r-ésima ocurrencia de E.

Entonces, la variable aleatoria:

$$X = \sum_{i=1}^{r} X_i$$

es el número de repeticiones independientes de un ensayo de Bernoullí hasta que se obtenga el éxito número r. Luego, X tiene distribución de Pascal con parámetros r y p.

TEOREMA 7.4.

Si X es una variable aleatoria con distribución de Pascal con parámetros r y p, entonces,

$$\text{a) } \mu = r\frac{1}{p}, \qquad \text{b) } \sigma^2 = r\frac{q}{p^2}.$$

PRUEBA.

Sean $X_1, X_2, ..., X_r$, r variables aleatorias independientes cada una con distribución geométrica, $G(p)$, entonces, la variable aleatoria:

$$X = \sum_{i=1}^{r} X_i$$

tiene distribución de Pascal con parámetros r y p. Luego,

$$E(X) = E\left(\sum_{i=1}^{r} X_i\right) = \sum_{i=1}^{r} E(X_i) = \sum_{i=1}^{r} \frac{1}{p} = r\frac{1}{p}$$

$$V(X) = V\left(\sum_{i=1}^{r} X_i\right) = \sum_{i=1}^{r} V(X_i) = \sum_{i=1}^{r} \frac{q}{p^2} = r\frac{q}{p^2}.$$

EJEMPLO 7.6

Una máquina produce artículos de uno en uno y de manera independiente. Se considera que el 10% de ellos son defectuosos. Si la máquina se detiene apenas produce el cuarto artículo defectuoso:

a) ¿Cuál es el número esperado de artículos producidos hasta que se detiene la máquina?.
b) ¿Cuál es la probabilidad de que la máquina se detenga en el décimo artículo producido?.
c) ¿Cuál es la probabilidad de que produzca al menos 10 artículos para que la máquina se detenga?.

SOLUCION

Sea X el número de artículos producidos hasta tener 4 defectuosos. La ley de probabilidad de X es Pascal con parámetros $r = 4$ y $p = 0.1$. Esto es,

$$P[Y = k] = C_3^{k-1} p^4 (1-p)^{k-4}, \qquad k = 4, 5, 6, \ldots, etc.$$

a) $E(X) = r\dfrac{1}{p} = 4 \times \dfrac{1}{0.1} = 40$

b) $P[X = 10] = C_3^9 p^4 (1-p)^6$

c) $P[X \geq 10] = 1 - P[4 \leq X \leq 9] = 1 - \sum_{k=4}^{9} C_3^{k-1} p^4 (1-p)^{k-4}$.

7.1.5 Distribución hipergeométrica

Definición. Un **experimento hipergeométrico** consiste en escoger al azar una muestra de tamaño n , **uno a uno sin restitución**, de N elementos o resultados posibles, donde r de los cuales pueden clasificarse como éxitos, y los $N - r$ restantes como fracasos.

En cada extracción, la probabilidad de que el elemento sea un éxito es diferente, ya que la extracción es sin reposición.

Definición. Se denomina variable aleatoria hipergeométrica a la variable X que se define como el número de éxitos en el experimento hipergeométrico.

Si $n \leq r$, el rango de X es el conjunto $R_X = \{0,1,2,\ldots,n\}$.

Sea $k \in R_X$, entonces, el evento $[X = k]$ ocurre, si y sólo si en las n extracciones sucesivas sin reposición ocurren k éxitos y $n - k$ fracasos. El total de formas que puede ocurrir es:

$$C_k^n V_k^r V_{n-k}^{N-r}$$

Por otra parte, el número de formas diferentes de escoger n elementos de N una a una sin reposición es igual a V_n^N.

Luego, *la probabilidad de que ocurran k éxitos en el experimento hipergeométrico es*:

$$P[X=k]=\frac{C_k^n V_k^r V_{n-k}^{N-r}}{V_n^N}=\frac{C_k^r C_{n-k}^{N-r}}{C_n^N}.$$

Definición. Se dice que **la variable aleatoria X que se define como el número de éxitos en una muestra de tamaño n que se selecciona al azar uno por uno sin reposición** de N elementos o resultados posibles, de los cuales r son clasificados como éxitos y los restantes $N-r$ como fracasos, tiene **distribución hipergeométrica** y se escribe $X\sim H(N, n, r)$, si su función de probabilidad es

$$f(k)=P[X=k]=\frac{C_k^r C_{n-k}^{N-r}}{C_n^N}, \quad k=0,1,2,\dots,n.$$

Para probar que la suma de las probabilidades es igual a 1, se usa la identidad:

$$\sum_{k=0}^{n} C_k^r C_{n-k}^{N-r}=C_n^{r+N-r}=C_n^N.$$

NOTA. En el numerador de la función de probabilidad hipergeométrica, se requiere que $N-r\geq n-k$, de donde resulta, $k\geq n+r-N$, luego, el menor valor que toma la variable aleatoria X es el número:

$$a=\max(0, n+r-N).$$

Por otro lado, el mayor valor debe verificar $k\leq n$ y $k\leq r$, luego, el mayor valor que toma X puede denotarse por:

$$b=\min(n, r).$$

TEOREMA 7.5.

Sea X una variable aleatoria con distribución hipergeométrica $H(N,n,r)$, y sean, $p=r/N$, $q=1-p$, entonces,

a) $E(X)=np$

b) $Var(X)=npq\dfrac{N-n}{N-1}$

c) $H(N,n,r) \cong B(n,p)$. Esto es, si N tiende a $+\infty$ o N es grande con respecto a n, entonces **la distribución hipergeométrica $H(N, n, r)$ se aproxima a una distribución binomial $B(n,p)$.**

PRUEBA.

Utilizando la identidad $C_n^N = \frac{N}{n} C_{n-1}^{N-1}$, se tiene:

a) $$E(X) = \sum_{k=0}^{n} k \frac{C_k^r C_{n-k}^{N-r}}{C_n^N} = r.\frac{n}{N}.\frac{1}{C_{n-1}^{N-1}}.\sum_{k=1}^{n} C_{k-1}^{r-1} C_{n-k}^{N-r} = n \times \frac{r}{N}.$$

b) Se verifica $E[X(X-1)] = n(n-1) \times \frac{r}{N} \times \frac{r-1}{N-1}$, entonces,

$$Var(X) = E\left(X^2\right) - \mu^2 = E[X(X-1)] + \mu - \mu^2$$

$$Var(X) = n(n-1) \times \frac{r}{N} \times \frac{r-1}{N-1} + n.\frac{r}{N} - \left(n \times \frac{r}{N}\right)^2$$

$$= n \times \frac{r}{N}\left[(n-1) \times \frac{r-1}{N-1} + 1 - \frac{nr}{N}\right] = n \times \frac{r}{N}\left[\frac{(N-r)(N-n)}{N(N-1)}\right]$$

$$Var(X) = n \times \frac{r}{N} \times \left(1 - \frac{r}{N}\right) \times \left(\frac{N-n}{N-1}\right).$$

c) Aquí, N tiende a $+\infty$ se interpreta como: **"si N es grande con respecto a n".** Si N es grande con respecto a n, entonces, se pueden hacer las siguientes aproximaciones:

$$V_n^N = N(N-1)(N-2)...(N-n+1) \cong N^n$$

$$V_k^r = r(r-1)(r-2)...(r-k+1) \cong r^k$$

$$V_{n-k}^{N-r} = (N-r)(N-r+1)...(N-r-n+k+1) \cong (N-r)^{n-k}$$

Luego, si N es infinitamente grande con respecto a n, se tiene:

$$P[X=k] = \frac{C_k^n V_k^r V_{n-k}^{N-r}}{V_n^N} \cong C_k^n.\frac{r^k (N-r)^{n-k}}{N^n} = C_k^n \left(\frac{r}{N}\right)^k \left(1 - \frac{r}{N}\right)^{n-k}$$

La aproximación es buena si $n \leq 0.1 \times N$.

Observe que la media de la binomial y la hipergeométrica es la misma, mientras que la varianza difiere en el factor de corrección $(N-n)/(N-1)$, que se reduce a uno cuando N es grande respecto a n.

EJEMPLO 7.7

Se escriben los nombres de 7 mujeres y 3 hombres en pedacitos de papel y se los coloca en una caja. De la caja se escogen al azar 4 papelitos.

a) Determine la distribución de probabilidad del número de papelitos seleccionados que contengan nombres de hombres:
   a1) Si se escogen uno por uno sin reposición
   a2) si se escogen uno por uno con reposición

b) Calcule la probabilidad de que se seleccionen los nombres de por lo menos dos hombres.

SOLUCION.

a) Sea X la variable aleatoria que se define como el número de papelitos que contengan nombres de hombres de los 4 seleccionados:

a1) Si se escogen uno por uno sin reposición, los valores posibles de X son: 0,1,2,3. Entonces, $X \sim H(10,4,3)$, y la función de probabilidades de X es:

$$P[X=x]=\frac{C_x^3 C_{4-x}^7}{C_4^{10}}, \qquad x=0,1,2,3$$

a2) Si se escogen uno por uno con reposición, los valores posibles de X son 0,1,2,3,4. Entonces, $X \sim B(4,p)$, con $p=3/10=0.3$ y la función de probabilidades de X es:

$$P[X=x]=C_x^4 p^k (1-p)^{4-k}, \qquad x=0,1,2,3,4$$

b) La probabilidad de que se seleccionen los nombres de por lo menos dos hombres

$$P[X \geq 2]=1-P[0 \leq X \leq 1]$$

Si se escogen uno por uno sin reposición se tiene:

$$P[X \geq 2]=1-\frac{C_0^3 C_4^7 + C_1^3 C_3^7}{C_4^{10}}=1-\frac{140}{210}=0.333.$$

Si se escogen uno por uno con reposición se tiene:

$$P[X \geq 2]==1-[C_0^4 (0.3)^0 (0.7)^4 + C_1^4 (0.3)^1 (0.7)^3]=1-0.652=0.348$$

EJEMPLO 7.8

Una compañía recibe semanalmente un embarque de 500 artículos de cierto tipo. La Cia. controla la calidad de cada embarque probando 10 artículos escogidos al azar uno por uno sin reposición y rechaza el embarque si más de uno de los artículos probados no cumplen las especificaciones. Se sabe que cada embarque semanal contiene 90% de artículos que cumplen las especificaciones.

Sea X el número de artículos en la muestra que no cumplen las especificaciones.

a) ¿Con qué probabilidad se rechaza cualquier embarque semanal?

b) Si el costo de la inspección semanal en soles está dado por : $C = 2 + 4X + X^2$ calcular el costo esperado por inspección

c) Hallar la probabilidad de que el costo de inspección sea mayor o igual que 34 soles. Resolver usando el modelo hipergeométrico y la aproximación binomial.

SOLUCION.

Los posibles valores de X son: 0,1,2,...,10 y su distribución es hipergeométrica $H(500,50,10)$, esto es:

$$P[X = k] = C_k^{50} C_{10-k}^{450} \Big/ C_{10}^{500}, \quad k = 0,1,\ldots,10$$

a) La probabilidad de rechazar cualquier embarque semanal es igual a:

$$P[X \geq 2] = 1 - P[X \leq 1] = 1 - \sum_{k=0}^{1} C_k^{50} C_{10-k}^{450} \Big/ C_{10}^{500} = 1 - 0.7365 = 0.2635$$

b) $E(C) = 2 + 4E(X) + E(X^2) = 2 + 4 \times 1 + 1.81 = 7.81$

donde $E(X) = 10 \times \dfrac{50}{500} = 1$, $\sigma^2 = 10 \times 0.1 \times 0.9 \times \dfrac{500-50}{500-1} = 0.81$

$$E(X^2) = \sigma^2 + \mu^2 = 0.81 + (1)^2 = 1.81$$

c) $P[C \geq 34] = P[X \geq 4] = 1 - P[X \leq 3] = 1 - \sum_{k=0}^{3} C_k^{50} C_{10-k}^{450} \Big/ C_{10}^{500} = 0.01186$

Aproximando la hipergeométrica $H(500,50,10)$ a la binomial $B(10,0.1)$, se tiene:

$$P[X \geq 4] = \sum_{x=4}^{10} C_x^{50} C_{10-x}^{450} \Big/ C_{10}^{500} \cong \sum_{x=4}^{10} C_x^{10} (0.1)^x (0.9)^{10-x} = 0.0128.$$

7.1.6 Distribución de Poisson

Definición. Se dice que la variable aleatoria discreta X, cuyos valores posibles son: 0,1,2,..., tiene distribución de Poisson con parámetro λ $(\lambda > 0)$ y se escribe $X \sim P(\lambda)$, si su función de probabilidad es:

$$f(x) = P[X = x] = \frac{e^{-\lambda}(\lambda)^x}{x!}, \qquad x = 0,1,2,...$$

Para verificar que la suma de las probabilidades de Poisson es igual a 1, se utiliza la suma: $\displaystyle\sum_{k=0}^{\infty}\frac{(\lambda)^k}{k!} = e^{\lambda}$.

En efecto, $$\sum_{k=0}^{\infty}\frac{e^{-\lambda}(\lambda)^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda)^k}{k!} = e^{-\lambda}e^{\lambda} = e^0 = 1.$$

NOTA. **La distribución de Poisson se aplica** a problemas donde la variable aleatoria es el número de eventos independientes que ocurren en un intervalo de tiempo, o en una región plana (con un promedio dado), por ejemplo, entre otros:

* Número de llamadas que recibe una central telefónica en el período de un minuto.
* Número de accidentes de trabajo que ocurren en una fábrica durante una semana.
* Número de fallas en la superficie de una cerámica rectangular.
* Número de bacterias en un volumen de un mt^3 de agua.

TEOREMA 7.6

Si $X \sim P(\lambda)$, entonces, a) Media: $\mu = \lambda$, b) Varianza: $\sigma^2 = \lambda$

PRUEBA.

a) $$E(X) = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}(\lambda)^x}{x!} = \sum_{x=1}^{\infty} x\frac{e^{-\lambda}(\lambda)^x}{x!} = (\lambda)e^{-\lambda}\sum_{x=1}^{\infty}\frac{(\lambda)^{x-1}}{(x-1)!} = (\lambda)e^{-\lambda}e^{\lambda} = \lambda$$

b) $$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1).\frac{e^{-\lambda}(\lambda)^x}{x!} = \lambda^2.e^{-\lambda}.\sum_{x=2}^{\infty}\frac{(\lambda)^{x-2}}{(x-2)!} = \lambda^2.e^{-\lambda}.e^{\lambda} = \lambda^2$$

Como $$Var(X) = E[X(X-1)] + \mu - \mu^2,$$

entonces, $$\sigma^2 = E[X(X-1)] + \mu - \mu^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

EJEMPLO 7.9.

Suponga que llegan en forma aleatoria una serie de llamadas a una central telefónica con un promedio de tres llamadas por minuto.

a) Calcular la probabilidad de que en el periodo de un minuto

a1) no ocurra llamada alguna, a2) ocurran al menos 4 llamadas

b) Si cada llamada cuesta S/.0.50, ¿cuánto es el costo esperado por llamada?

SOLUCION.

Sea X el número de llamadas que ocurren en el período de un minuto. $X \sim P(\lambda)$, donde $\lambda = 3$ es el promedio del número de llamadas por minuto.

a) La probabilidad de que ocurran k llamadas en el período de un minuto es:

$$P[X = k] = \frac{e^{-\lambda}(\lambda)^k}{k!} = \frac{e^{-3}(3)^k}{k!}, \qquad k = 0,1,2,\ldots, \text{etc.}$$

a1) La probabilidad de que no ocurra llamada alguna en el periodo de un minuto es:

$$P[X = 0] = \frac{e^{-3}(3)^0}{0!} = 0.0498.$$

a2) La probabilidad de que ocurran al menos 4 llamadas en el periodo de un minuto es :

$$P[X \geq 4] = 1 - P[X \leq 3] = 1 - \sum_{k=0}^{3} \frac{e^{-3}3^k}{k!} = 1 - 0.64723 = 0.35277.$$

b) Sea C el costo por llamada, entonces, $C = 0.5X$, y

$$E(C) = 0.5E(X) = 0.5 \times 3 = 1.5$$

NOTA. (Extensión o reducción del intervalo unitario)

La probabilidad de que ocurran k eventos de Poisson **en un intervalo de tiempo o en una región de tamaño t** es

$$P[X = k] = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \qquad k = 0,1,2,\ldots, etc$$

donde λ es el número promedio de ocurrencias **por unidad de período o región** (λt **es el número promedio** de ocurrencias de eventos en el período o región de tamaño t)

EJEMPLO 7.10.

Una empresa textil produce un tipo de tela en rollos de 100 metros. El número de defectos que se encuentra al desenrollar la tela es una variable aleatoria de Poisson que tiene en promedio 4 defectos por cada 20 metros de tela.

a) ¿Qué probabilidad hay de que al desenrollar la tela se encuentre menos de tres defectos en los primeros 50 metros?.
b) Hallar la probabilidad de que al desenrollar la tela no se encuentre defectos en el primer segmento de 5 metros de tela.
c) Si se desenrollan 5 rollos de la tela escogidos al azar, ¿cuál la probabilidad de que no se encuentre defectos en el primer segmento de 5 metros de tela en al menos dos de ellos?.

SOLUCION.

Sea X el número de defectos encontrados en un segmento de 20 metros de tela que ocurre con promedio $\lambda = 4$. La probabilidad de encontrar k defectos en el segmento de $20 \times t$ metros de tela es:

$$P[X = k] = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \qquad k = 0,1,2,..,etc.$$

donde λt es el promedio de defectos en el segmento de $20 \times t$ metros de tela,

a) El promedio de defectos en los primeros 50 metros de tela es $\lambda t = 4 \times 2.5 = 10$ ($t = 2.5$ aumenta la longitud de 20 a 50 metros) y la probabilidad de que se encuentren menos de tres defectos en los primeros 50 metros de tela es:

$$P[X < 3] = P[X \leq 2] = \frac{e^{-10}(10)^0}{0!} + \frac{e^{-10}(10)^1}{1!} + \frac{e^{-10}(10)^2}{2!} = 0.00277$$

b) El promedio de defectos en los primeros 5 metros de tela es $\lambda t = 4(1/4) = 1$ ($t = 1/4$ reduce la longitud de 20 a 5) y la probabilidad de no encontrar defectos en los primeros 5 metros de tela es

$$P[X = 0] = \frac{e^{-1}(1)^0}{0!} = 0.3679$$

c) Sea Y el número de rollos que no tienen defectos en el primer segmento de 5 metros de tela de 5 rollos de tela escogidos al azar. La variable Y cuyos valores posibles son: 0, 1, 2, 3, 4, 5 se distribuye según la ley binomial $B(5, p)$, donde $p = P[X = 0] = 0.3679$. Luego,

$$P[Y \geq 2] = 1 - P[Y \leq 1] = 1 - \sum_{k=0}^{1} C_k^5 p^k (1-p)^{5-k} = 1 - 0.3946 = 0.6054.$$

Aproximación de la distribución Binomial a la Poisson

TEOREMA 7.7. Sea X una variable aleatoria con distribución binomial $B(n,p)$. Si cuando $n \to +\infty$, $p \to 0$, y $\lambda = np$, **permanece constante,** entonces, la distribución binomial se aproxima a la distribución de Poisson con parámetro λ. esto es,

$$P[X=x] = C_x^n p^x (1-p)^{n-x} \quad \text{tiende a} \quad P[X=x] = \frac{e^{-\lambda}\lambda^x}{x!}.$$

PRUEBA.

$$P[X=x] = C_x^n p^x (1-p)^{n-x} = \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!}.p^x(1-p)^{n-x}$$

Sustituyendo $p = \dfrac{\lambda}{n}$, se tiene

$$P[X=x] = \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$P[X=x] = \frac{\lambda^x}{x!}\left[(1)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)\right]\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

Por otro lado, si cuando $n \to +\infty$, $np = \lambda$ permanece constante, resulta:

$$\lim_{n \to +\infty}\left(1-\frac{i}{n}\right) = 1, \text{ para } \quad i = 1,2,3,\ldots(x-1)$$

$$\lim_{n \to +\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}, \qquad \lim_{n \to +\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1$$

Luego, $\displaystyle\lim_{n \to +\infty} C_x^n p^x (1-p)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}.$

La aproximación binomial a Poisson es aceptable si $n \geq 30$ y $p \leq 0.05$.

EJEMPLO 7.11.

Un estudio realizado en las tierras de cultivo de Tarapoto concluye afirmando que la probabilidad de que cada hectárea de siembra de arroz contenga por lo menos un nido de hormiga es de 0.005. De 600 hectáreas de siembras de arroz escogidas al azar, ¿qué probabilidad hay de que al menos 5 de ellas contengan por lo menos un nido de hormiga?.

SOLUCION.

Sea X el número hectáreas, de las 600 de siembras de arroz escogidas, que contengan por lo menos un nido de hormiga. Los posibles valores de X son 0,1,2,...,600 y se distribuyen según el modelo binomial: $B(600, 0.005)$, esto es:

$$P[X = k] = C_k^{600}(0.005)^k (0.995)^{600-k}. \quad k = 0, 1, 2, 3, ..., 600$$

La probabilidad de que al menos 5 de las hectáreas de siembras de arroz contengan por lo menos un nido de hormiga es:

$$P[X \geq 5] = 1 - P[X \leq 4] = 1 - \sum_{k=0}^{4} C_k^{600}(0.005)^k (0.995)^{600-k}$$

Si utilizamos la **aproximación a la Poisson**, se tiene: $\lambda = np = 600(0.005) = 3$

$$C_k^{600}(0.005)^k (0.995)^{600-k} \cong \frac{e^{-3}(3)^k}{k!} \quad \text{y}$$

$$P[X \geq 5] = 1 - P[X \leq 4] \cong 1 - \sum_{k=0}^{4} \frac{e^{-3}3^k}{k!} = 1 - 0.81526 = 0.18474$$

EJERCICIO.

El número de accidentes que sufre una empresa de Taxis se distribuye según el modelo de Poisson con un promedio de 2 accidentes por semana.

a) Si el primer día de la semana ya ha habido un accidente, ¿cuál es la probabilidad de que en esa semana no haya más de dos accidentes?

b) ¿Cuántos autos tiene la Cia. de Taxis si existe una probabilidad igual a 0.25 de que no hay taxi que no sufra accidente en una semana?

SOLUCION

a) Sea X: # de accidentes que sufre la Cia. en el periodo de una semana. X tiene distribución de Poisson con promedio λ=2 por semana. Luego,

$$P[X \leq 2 / X \geq 1] = \frac{P[1 \leq X \leq 2]}{P[X \geq 1]} = \frac{\frac{e^{-2}2^1}{1!} + \frac{e^{-2}2^2}{2!}}{1 - \frac{e^{-2}2^0}{0!}} = 0.626$$

b) Sea Y: # de taxis que no sufren accidentes en una semana, de los n taxis que tiene la Cia. $Y \sim B(n, p)$, donde $p = P[X = 0] = e^{-2}$

Se debe hallar n tal que, $P[Y = 0] = C_0^n p^0 q^n = 0.25$, entonces, $(1 - e^{-2})^n = 0.25$,

de donde resulta $$n = \frac{\log(0.25)}{\log(1 - e^{-2})} = 9.53 \cong 10$$

EJERCICIOS

Binomial

1. Si $X \sim B(n, p)$ ta. que $E(X) = 3$ y $V(X) = 2.4$, calcular: $P[X \geq 3]$.

Rp. n =15, p=0.2, Prob=0.60198.

2. Un estudiante contesta al azar (o sea sin saber nada) a 9 preguntas, siendo cada una de 4 respuestas de las cuales sólo una es la correcta.

a) Determinar la distribución de probabilidades del número de preguntas contestadas correctamente.

b) Si para aprobar tal examen debe contestar correctamente al menos 6 preguntas, ¿cuál es la probabilidad de aprobar el examen?.

Rp. a) $B(9, 0.25)$. b) 0.01.

3. En una producción, la probabilidad de que un objeto sea defectuoso es 0.2. Si en una muestra de n de tales objetos escogidos al azar uno por uno, se espera que haya un defectuoso,

a) calcular la probabilidad de que haya un objeto defectuoso.

b) ¿cuántos objetos defectuosos es más probable que ocurra?

Rp. a) $n = 5$, 0.4096, b) 1.

4. El 75% de la mercadería que recibe un comerciante del fabricante A es de calidad excepcional, mientras que el 80% de la mercadería que recibe del fabricante B es de calidad excepcional. El 60% del total de la mercadería lo adquiere de A y el resto de B. Si se seleccionan 4 unidades de la mercadería, ¿qué probabilidad hay de que se encuentren 2 unidades que sean de calidad excepcional?.

Rp. p=0.77, $X \sim B(4,p)$, $P[X=2]$=0.188

5. En una empresa donde los empleados son 80% hombres y 20% mujeres están aptos para jubilarse el 10% de las mujeres y el 10% de los hombres. De cinco solicitudes para jubilarse, ¿cuál es la probabilidad de que al menos dos estén aptos para jubilarse?.

Rp. p=0.8×0.1+0.2×0.1=0.1, $X \sim B(5, 0.1)$, $P[X \geq 2]$=0.0815.

6. La producción de cuatro máquinas es recogida en cajas de 5 unidades. La experiencia permitió establecer la siguiente distribución de las cajas, según el número de unidades defectuosas que contienen;

# de unidades defectuosas	0	1	2	3	4	5
Porcentaje de cajas	0.70	0.15	0.08	0.05	0.02	0.00

La inspección diaria consiste en examinar las 5 unidades de cada caja.

Se acepta una caja, cuando contiene menos de dos unidades defectuosas. En caso contrario se rechaza.

a) ¿Cuál es la probabilidad de rechazar una caja que no contenga unidades defectuosas?
b) ¿Cuál es la probabilidad de aceptar una caja que contenga tres unidades defectuosas?

Rp. Sea X :# unidades defectuosas en la caja, a) $X \sim B(5, 0.7)$, $P[X \geq 2]=1-0.031=0.969$, b) $X \sim B(5, 0.05)$. $P[X \leq 1]=0.9774$.

7. Un vendedor a domicilio compra diariamente 10 unidades de un producto a \$2 cada una. Por cada producto gana 13\$ si lo vende o pierde 1\$ además del costo si no lo vende en el día. Si la probabilidad de venta de cada unidad es 0.2 y si las ventas son independientes.

a) hallar la distribución de probabilidad de las unidades vendidas.
b) calcular la utilidad esperada del vendedor.

Rp. a) $B(10, 0.2)$. b) \$2.

8. Una computadora utilizada por un sistema bancario de 24 horas asigna cada transacción al azar y con igual probabilidad, a una de cinco posiciones de memoria: 1, 2, 3, 4, 5. Si al terminar el periodo nocturno de un día se han registrado 15 transacciones, ¿cuál es la probabilidad de que el número de transacciones efectuadas a las posiciones de memoria par sea mayor que 3?

Rp. Sea X:# de éxitos de 15, éxito=transacc. pares, P(éxito)=2/5, $X \sim B(15,0.4)$, $P[X>3]=0.909$.

9. Una secretaria que debe llegar a la oficina a las 8 de la mañana, se retrasa 15 minutos en el 20% de las veces. El gerente de la compañía que no llega si no hasta las 10 de la mañana llama ocasionalmente a la oficina entre las 8 y 8.15 de la mañana para dictar una carta. Calcular la probabilidad de que en 5 mañanas por lo menos en una no encuentre a la secretaria.

Rp. Exito=No está de 8am a 8.15am. X: # de éxitos de $n=5$, $X \sim B(5,0.2)$, $P[X \geq 1]=0.672$.

10. Al realizar un experimento, la probabilidad de lograr el objetivo es 0.4. Si se realiza el experimento 20 veces bajo las mismas condiciones y asumiendo resultados independientes

a) Calcular la probabilidad de lograr el objetivo por lo menos en tres de las 20 veces
b) El costo de realizar el experimento es de S/.1500, si se logra el objetivo; y de S/.3000 si no se logra. Calcular el costo esperado para realizar el experimento.

Rp. X:# éxitos de $n=20$. éxito=lograr el objetivo, $p=0.4$. $X \sim B(20,0.4)$, $E(X)=8$ a) $P[X \geq 3]=0.996$, b) costo=$1500X+3000(20-X)$, E(costo)=48000.

11. Una prueba de aptitud consta de 10 preguntas con cinco alternativas cada una, de las cuales sólo una es la correcta. La calificación se realiza de la siguiente manera: Cada pregunta correctamente contestada vale 2 puntos. Por cada pregunta mal contestada se descuenta k puntos.

a) Determine el valor de k de tal manera que la nota esperada de un alumno que responde al azar las 10 preguntas sea cero.

b) ¿Cuál es la probabilidad de que un alumno que responde al azar las 10 preguntas , tenga una nota mayor o igual a 11?.

Rp. a) X: # de respuestas correctas, $X \sim B(10, 0.2)$, $E(X)=2$, Nota$=2X-k(10-X)$, a) $E(\text{Nota})=0$, implica $k=0.5$, b) $P[\text{Nota}\geq 11]=P[X\geq 7]=0$.

12. Una agencia de viajes observa que el 20% de las reservas de pasajes para viajar al interior no se hacen efectivas. Si la compañía decide acepar reservas por un 20% más de los 15 cupos que posee, calcular el porcentaje de clientes que habiendo hechos sus reservas se quedarían sin viajar.

Rp. X: # de cupos cubiertos de 18. $X \sim B(18, 0.80)$, $P[X>15]=0.271$.

13. El tiempo de duración X, en meses, de un tipo de resistencia eléctrica tiene función de densidad de probabilidad:

$$f(x)=\begin{cases} 0.5e^{-0.5x}, & \text{si } x\geq 0 \\ 0, & \text{en otro caso} \end{cases}$$

a) ¿Cuál es la probabilidad de que una de tales resistencias eléctricas dure más de 4 meses?.

b) Si se prueban 10 resistencias eléctricas, ¿cuál es la probabilidad de que ninguna dure más de 4 meses?.

c) ¿Cuántas resistencias se probarían para que con probabilidad igual a 0.9 se tenga al menos uno que dure más de 4 meses?

d) Si el costo de producción de una resistencia es:

$$C = 2 + (30 - X)^2$$

¿cuánto es el valor esperado del costo?

Rp. a) $p=P[X>4]=e^{-2}=0.135$, b) $Y \sim B(10,p)$, $P[Y=0]=(0.865)^{10}$, c) $W \sim B(n,p)$, $0.9=P[W\geq 1]$, implica, $(0.865)^n=0.1$, entonces, $n=15.877\cong 16$, d) $E(C)=790$.

Geométrica y Pascal

14. En cierto proceso de producción se sabe que el porcentaje de artículos defectuosos es de 0.02. Se controlan la calidad de los artículos uno por uno

a) Calcular la probabilidad de que el décimo artículo probado sea el primer defectuoso encontrado.

b) En promedio, ¿cuántos artículos se probarían hasta encontrar el primer defectuoso?

Rp. a) 0.017, b) 1/0.02= 50.

15. Cierto virus ha invadido un colegio atacando al 5% de los niños. Si tales niños son examinados uno por uno, ¿cuál es la probabilidad de que el doceavo niño examinado sea el quinto niño encontrado atacado por el virus?.

Rp. $C_4^{11}(0.05)^5(0.95)^7$.

16. Un vendedor a domicilio estima en 0.3 la probabilidad de efectuar una venta. Suponga ventas independientes.

a) En promedio, ¿cuántas visitas domiciliarias debe efectuar hasta tener 12 ventas?.

b) ¿Cuál es la probabilidad de que se hagan al menos 10 visitas para tener 4 ventas?.

Rp. a) 40, b) $1-\sum_{k=4}^{9} C_3^{k-1}(0.7)^{k-4}(0.3)^4$

17. En un proceso de producción, la probabilidad de que se produzca cada artículo que cumpla ciertas especificaciones es 0.99. En determinado momento se plantea el objetivo de producir 150 artículos que cumplan con las especificaciones; pero al mismo tiempo se decide detener el proceso de producción tan luego se produzca el primer artículo que no cumpla con las especificaciones.

a) ¿Cuál es la probabilidad de lograr el objetivo?.

b) Si después de producir 100 artículos aún no se ha detenido el proceso, ¿cuál sería la probabilidad de lograr el objetivo?.

Rp. X: # de artículos producidos hasta que ocurra 1er defectuoso. $X \sim G(0.01)$, $k=1,2,$ etc.,
a) $P[X>150]=(0.99)^{150}$, b) $P[X>150/X>100]=(0.99)^{50}$.

18. Una compañía petrolera ha sido designada para perforar pozos en la amazonía peruana hasta obtener un resultado exitoso. La Compañía estima en 0.7 la probabilidad de no hallar petróleo para cada pozo que perfora.

a) Suponga que la compañía petrolera cree que una serie de exploraciones será rentable si el número de pozos perforados hasta que ocurra el primer éxito es menor o igual que 5. Calcule la probabilidad de que la exploración no será rentable si ya fueron perforados 3 pozos y en ninguno de ellos se encontró petróleo.

b) El costo para perforar cada pozo es de $10.000. Si un ensayo no resulta exitoso, el siguiente ensayo tiene un costo adicional de $5000. ¿Cuánto es el costo esperado del proyecto?

c) Si la compañía dispone de un presupuesto de $145.000, ¿cuál es la probabilidad de que los trabajos experimentales tengan un costo que sobrepase el presupuesto de la compañía?.

Rp. X = # de perforaciones hasta obtener éxito. $X \sim G(p)$, $p=0.3$, a) $P[X>5/X>3]=P[X>2]=(0.7)^2$,
b) $C(X) = 15{,}000X - 5{,}000$, $E(C(X)) = \$45{,}000$, c) $P[C(X)>145{,}000]=P[X>10]=(0.7)^{10}$

19. Un experimento se repite de manera independiente hasta que se obtiene el primer éxito. El costo C producido, está en función del número X de repeticiones necesarias hasta tener éxito y es dado por:

$$C = 200 + 5X + 4X^2$$

Se sabe que la varianza del número de repeticiones necesarias hasta obtener éxito es 40/36. Si en estas condiciones, el experimento debe ser realizado por 200 personas hasta que cada una de ellas obtenga éxito, ¿cuánto sería el costo esperado?.

Rp. $E(C)$=200+5$E(X)$+4$E(X^2)$=221.2, de Var(X)=40/36=$(1-p)/p^2$, resulta p=0.6, $E(X)$=1/0.6, V(X)=[40/36+(10/6)^2], 200$E(C)$=200×221.2=42,443.8.

20. En un lote de 15 artículos existen n que son defectuosos. Se extraen uno por uno con restitución hasta obtener el primer defectuoso. El costo total por inspección es 2^X donde X es el número de artículos inspeccionados. Hallar n si el costo total esperado por inspección es $16.

Rp. $X\sim$ G(p), con $p=n/15$. De 16=$E(2^X)$=$2p/(2p-1)$, sale p=8/15, y $n = 8$.

21. Verificar que: $\displaystyle\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$

Rp. Derive respecto a q ambos lados de $\displaystyle\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$.

Hipergeométrica

22. Una caja contiene 20 chips, de los cuales el 20% no están aptos para su venta y el resto si. Se escogen 5 chips al azar uno a uno sin reposición.

a) Determine la distribución de probabilidad del número de chips escogidos que sean no aptos para su venta.

b) Calcular la probabilidad de que al menos uno sea no apto para su venta.

Rp. a) $P[X=k]$= $C_k^4 C_{5-k}^{16} / C_5^{20}$,k=0,1,2,3,4 , b) $P[X\geq 1]=1-P[X=0]=1-(C_5^{16}/C_5^{20})$.

23. Un determinado producto industrial es embarcado en lotes de 20 unidades. Se escogen 5 ítems al azar de un lote y se rechaza el lote si se encuentra dos o más defectuosos; en caso contrario se acepta el lote. Calcular la probabilidad de aceptar un lote que tiene 3 defectuosos si los ítems se escogen uno por uno,

a) con reposición, b) sin reposición.

Rp. a) $X\sim B(5, 0.15)$, $P[X\leq 1]$=0.8352. b) $X\sim H(20,3,5)$, $P[X\leq 1]$=0.8596.

24. Una ensambladora de computadoras recibe lotes de 10 tarjetas cada uno de un tipo específico, en la proporción, 30% de marca A y 70% de marca B. Se sabe

que el porcentaje de producción defectuosa es de 40% para la marca A y de 10% para la marca B. Si se prueban 3 tarjetas extraídas al azar una a una y sin reposición de un lote elegido al azar, calcular la probabilidad de no encontrar tarjetas defectuosas.

Rp. $\frac{3}{10}\frac{C_0^4 C_3^6}{C_3^{10}}+\frac{7}{10}\frac{C_0^1 C_3^9}{C_3^{10}}=0.54.$

25. Un jurado de 9 jueces va a decidir la inocencia o culpabilidad de un reo. Suponga que 6 votan por la inocencia y el resto por la culpabilidad. Si se seleccionan al azar 3 jueces y se pregunta por su voto ,

a) Calcular la media y la varianza del número de jueces de la selección que votan por la inocencia.

b) ¿Cuál es la probabilidad de que ningún juez seleccionado esté a favor de la inocencia del reo?.

Rp. a) $\mu=3(6/9)=2$, $\sigma^2=0.5$, b) 0.0119.

26. Un lote grande de N sacos de café de un quintal contiene 10% de sacos con impurezas. Del lote se cargan a un camión 20 sacos escogidos al azar.

a) Determine la distribución de probabilidad del número de sacos de café con impurezas escogidos, ¿cuántos sacos de café con impurezas se espera cargar al camión si se cargan muchas veces?.

b) Si N =1,000, calcular aproximadamente la probabilidad de que se haya obtenido a lo más 6 sacos con impurezas.

Rp. a) $H(N, 0.1N, 20)$, Esperado=2, b) Aprox a la binomial. Prob= 0.9976.

27. Una compañía recibe un envío de 20 piezas de un proceso de manufactura. El control de calidad de la Cía. consiste en tomar una muestra de 3 piezas al azar sin reposición de este envío. Si en la muestra se encuentra al menos una defectuosa, se rechaza el lote; en caso contrario se eligen al azar 2 piezas adicionales. Si en la segunda muestra se encuentra al menos una defectuosa, se rechaza el lote; en caso contrario se acepta el lote. Calcular la probabilidad de rechazar el envío si contiene 25% de piezas defectuosas.

Rp. 0.8.

28. En una fábrica la producción semanal de cierto tipo de artículo es de 1000 unidades. Cada semana se realiza un control de calidad seleccionando una muestra de 20 unidades del artículo, escogidos al azar y sin reposición de la producción de la semana (1000 artículos) y adoptando la siguiente regla de decisión: aceptar que el porcentaje de producción defectuosa es 2% si en la muestra se encuentra a lo más un defectuoso, o rechazar que el porcentaje de producción defectuosa es 2% en caso contrario, ¿cuál es la probabilidad de que

se haya decidido aceptar que el porcentaje de producción defectuosa de una semana es 2% cuando en realidad fue 5% de los 1000 producidos?

Rp. X:# defect en muestra de n=20. $X \sim H(1000, 50.20)$,

P[aceptar]=$P[X\leq 1/p=0.05]=(C_0^{50}C_{20}^{950}+C_1^{50}C_{19}^{950})/C_{20}^{1000}$

Poisson

29. El número medio de automóviles que llegan a una garita de peaje es de 120 por hora.

a) Calcular la probabilidad de que en un minuto cualquiera no llegue automóvil alguno.

b) Calcular la probabilidad de que en el período de 3 minutos lleguen más de 5 automóviles.

c) Si tal garita puede atender a un máximo de 3 automóviles en 30 segundos, calcular la probabilidad de que en un medio minuto dado lleguen más automóviles de lo que puede atender.

Rp. a) λ=2, e^{-2}=0.135. b) λt=6, $P[X>5]$=0.554. c) λt=1. $P[X>3]$= 0.0189..

30. Un líquido contiene cierta bacteria con un promedio de 3 bacterias por cm^3., calcular la probabilidad de que una muestra,

a) de 1/3 de cm^3. no contenga bacteria alguna.

b) de 2 cm^3. contenga por lo menos una bacteria.

Rp. λ=3, a) λt = 1, prob=0.368, b) λt=6, prob=0.998.

31. Suponga que el número de accidentes de trabajo que se producen por semana en una fábrica sigue la ley de Poisson de manera que la probabilidad de que ocurran 2 accidentes es igual a 3/2 de la probabilidad de que ocurra un accidente. Calcular la probabilidad de que no ocurran accidentes en 2 semanas consecutivas. Rp. λ=3, 0.002.

32. Un banco atiende todos los días de 8am. a 4pm. y se sabe que el número de clientes por día que van a solicitar un préstamo por más de $10,000 tiene una distribución de Poisson con una media de 3 clientes por día.

a) ¿Cuál es la probabilidad de que hasta el mediodía no se haya producido una solicitud de préstamo por más de $10,000?

b) En cuatro días, ¿cuál es la probabilidad de que en dos de los días hasta el mediodía no se haya producido una solicitud de préstamo por más de $10,000?

Rp. λ=3. a) λt=1.5, $P[X=0]=e^{-1.5}$=0.223, b) $C_2^4\,(0.223)^2(0.777)^2$.

33. La demanda semanal de cierto producto tiene una distribución de Poisson. Actualmente su media es 3 por semana. Se estima que después de una campaña

publicitaria, el valor esperado de la demanda se duplicará con probabilidad 0.8 y se triplicará con probabilidad 0.2. ¿cuál es la probabilidad de que después de la campaña la demanda sea igual a 4?.

Rp. P(demanda=4 después de la campaña)=$0.8e^{-6}(6)^4/4! + 0.2e^{-9}(9)^4/9!$

34. El número de personas que cada día se aloja en un hotel es una variable aleatoria X que tiene distribución de Poisson con parámetro λ, que puede ser:

igual a 20 con probabilidad 0.5,
igual a 15 con probabilidad 0.3,
igual a 10 con probabilidad 0.2

a) Determinar $P[X = k]$, donde $k = 1, 2, \ldots$, etc..
b) Calcular el valor esperado de X.

Rp. a) $P[X=k]= 0.5\,P[X=k/\lambda=20] + 0.3\,P[X=k/\lambda=15] + 0.2\,P[X=k/\lambda=10]$,
b) $E(X)=0.5\times20+0.3\times15+0.2\times10=16.5$.

35. Cierto tipo de loceta puede tener un número X de puntos defectuosos que sigue una distribución de Poisson con una media de 3 puntos defectuosos por loceta. El precio de la loceta es \$1 si $X = 0$, de \$0.70 si X=1 o 2, y de \$0.1 si X>2. Calcular el precio esperado por loceta.

Rp. X=0, prob=0.04979, X=1 o 2, prob=0,3734. X>2, prob=0.57681. esperado=\$0.37

36.. El número de usuarios que acuden a cierta base de datos confidencial sigue una distribución de Poisson con una media de dos usuarios por hora.

a) Calcular la probabilidad de que entre las 8 am. y el mediodía (12.m) acudan más de dos usuarios.
b) Si un operador de la base de datos trabaja todos los días de 8 am. hasta el mediodía (12.m), ¿cuál es la probabilidad de que este operador tenga que esperar más de 7 días hasta observar el primer día en el cual acceden más de dos usuarios?.

Rp. a) X: Poisson con λt=8. $P[X>2]$=0.986. b) Y~G(p), donde p=0.986. $P[Y>7]=(0.014)^7$

37. Cierta panadería dispone de una masa con frutas confitadas para hacer 200 panetones. Agrega 2,000 pasas de uvas a la masa y la mezcla bien. Suponga que el número de pasas es una variable aleatoria de Poisson con un promedio 10 pasas por panetón.

a) Calcular la probabilidad de que un panetón elegido al azar no contenga ninguna pasa.
b) ¿Cuántos panetones se espera que contengan 6 pasas?
c) Suponga que en tal producción hay 15 panetones con a lo más 6 pasas, si un cliente adquiere 5 panetones, ¿cuál es la probabilidad de que dos tengan mas de 6 pasas?.

Rp. a) e^{-10}, b) 200×$P[X=6]$=200×0.108=21.62≅22, c) $C_3^{15}C_2^{185}/C_5^{200}$.

38. Una compañía de seguros encuentra que el 0.1% de los habitantes de una gran ciudad fallece cada año en accidentes de tránsito. Calcular la probabilidad que la compañía tenga que pagar en un año a más de 10 de sus 3,000 asegurados contra tales accidentes.

Rp. $P[X>10]\cong 1-\sum_{k=0}^{10} e^{-\mu}\frac{\mu^k}{k!}=1-0.99971=0.00029$, donde $\mu=3000\times 0.001=3$

39. Suponga que la probabilidad de que se haga una soldadura defectuosa en un conexión dada es 0.001. Calcular la probabilidad de que se presenten a lo más 2 defectos en un sistema que tiene 5,000 conexiones soldadas independientemente.

Rp. $X\sim B(5000, 0.001)$ se aproxima a $P(\lambda)$ con $\lambda=5000\times 0.001=5$, $P[X\le 2]=0.125$

40. Un libro de n páginas contiene en promedio λ errores de impresión por página. Calcular la probabilidad que por lo menos una página contenga por lo menos $k+1$ errores.

Rp. Sea X: # de errores por página, $X\sim P(\lambda)$. Sea Y :# de páginas con por lo menos $k+1$ errores, $Y\sim B(n, p)$, $p=P[X\ge k+1]$ es la probabilidad de $k+1$ errores por página. La probabilidad que por lo menos una página contenga por lo menos $k+1$ errores es $P[Y\ge 1]=1-P[Y=0]$.

7.2 ALGUNAS DISTRIBUCIONES IMPORTANTES DE VARIABLES ALEATORIAS CONTINUAS

7.2.1 Distribución uniforme

Definición. Se dice que la variable aleatoria continua, X, tiene **distribución uniforme (o rectangular)** en el intervalo $[a, b]$, $a < b$, y se describe por $X \sim U[a,b]$, si su función de densidad de probabilidad es:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{si } a \le x \le b \\ 0, & \text{en otros casos} \end{cases}$$

La **función de distribución** del modelo uniforme es:

$$F(x) = \begin{cases} 0, & \text{si } x < a \\ \dfrac{x-a}{b-a}, & \text{si } a \le x < b \\ 1, & \text{si } x \ge b \end{cases}$$

La gráfica de la función de densidad y de la distribución acumulada se dan en la figura 7.1.

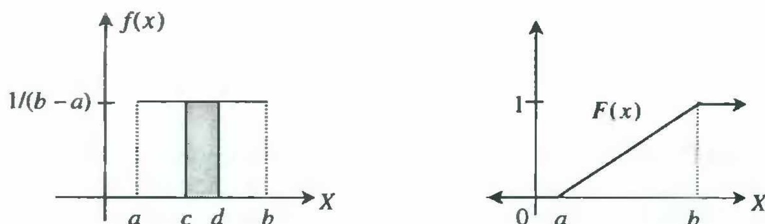


Figura 7.1 Función de densidad y de distribución acumulada uniforme

NOTA. Si $[c,d] \subset [a,b]$, la probabilidad de que X tome valores en el intervalo $[c, d]$ es:

$$P[c \le X \le d] = \frac{d-c}{b-a}$$

Esta probabilidad sólo depende de la longitud del intervalo $[c, d]$ y no de su ubicación dentro del intervalo $[a, b]$, y es la misma para todos los intervalos que tengan la mima longitud: $d-c$.

Algunas variables aleatorias continuas tienen distribución uniforme, por ejemplo:

- las coordenadas de un punto que se elige al azar en un intervalo dado.
- los instantes aleatorios en un período de tiempo dado.
- etc

TEOREMA 7.8.

Si la variable aleatoria X tiene distribución uniforme en el intervalo $[a,b]$, entonces,

a) $E(X)=\dfrac{a+b}{2}$ b) $Var(X)=\dfrac{(b-a)^2}{12}$.

EJEMPLO 7.12

Dos gerentes A y B deben encontrarse en cierto lugar entre las 7 p.m. y 8 p.m. para firmar un contrato. Cada uno espera al otro a lo más 10 minutos, ¿cuál es la probabilidad de que no se encuentren sabiendo que A llega a las 7.30 p.m.?.

SOLUCION.

Sea la variable aleatoria X el tiempo de llegada de B, que puede hacerlo en cualquier instante aleatorio entre las 7 p.m. y las 8 p.m. o entre 0 y 60 minutos. Entonces, $X \sim U[0,60]$ y su función de densidad de probabilidad es:

$$f(x)=\begin{cases}\dfrac{1}{60}, & \text{si } 0\le x\le 60\\ 0, & \text{en otros casos}\end{cases}$$

Puesto que A llega a las 7.30 p.m. o a los 30 minutos después de las 7 p.m. y espera a lo más 10 minutos, B no se encontrará con A si B llega de 7 p.m. a menos de 7.20 p.m. o si llega de después de las 7.40 p.m.. Luego, la probabilidad de que A y B no se encuentres es

$$P[0\le X<20 \text{ o } 40<X\le 60]=\int_0^{20}\frac{1}{60}dx+\int_{40}^{60}\frac{1}{60}dx=\frac{20}{60}+\frac{20}{60}=\frac{2}{3}$$

EJEMPLO 7.13

Un vendedor cobra honorarios fijos de S/.200 más una comisión de 5% del total de las ventas que realiza. Si el total de las ventas que realiza es una variable aleatoria X con distribución uniforme entre 0 y 2000 soles.

a) ¿Cuánto es la utilidad promedio del vendedor?

b) ¿Qué probabilidad hay de que obtenga honorarios superiores a S/.275?. ¿Cuánto debe vender como mínimo?.
c) Si vende como mínimo S/.500, ¿qué probabilidad hay de que gane más de S/.260?.
d) Se le propone al vendedor como honorarios el 32% de las ventas que realiza. Si quiere ganar por lo menos S/.240, ¿le conviene?
e) ¿Cuál es la probabilidad de que obtenga una ganancia igual a S/.250 ?

SOLUCION.

La utilidad o ganancia del vendedor es: $U = 200 + 0.05X$,donde, $X \sim U[0, 2000]$

a) $E(U) = 200 + 0.05 \times E(X) = 200 + 0.05 \times 1000 = 250$

b) $P[U > 275] = P[X > 1500] = \dfrac{500}{2000} = 0.25$. Debe vender mínimo S/.1500

c) $P[U > 260 / X > 500] = P[X > 1200 / X > 500] = \dfrac{800}{1500} = 8/15$

d) $P[0.32X > 240] = P[X > 750] = 1250/2000$, pero,

$P[200 + 0.05X > 240] = P[X > 800] = 1200/2000$. Si le conviene.

e) $P[U = 250] = P[X = 1000] = 0$

7.2.2 La distribución normal

Definición. Se dice que la variable aleatoria continua, X, que toma los valores reales, $-\infty < x < +\infty$, se **distribuye normalmente** (o más brevemente *es normal*) con parámetros μ y σ y se describe por $X \sim N(\mu, \sigma^2)$, si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \qquad -\infty < x < \infty .$$

donde $-\infty < \mu < \infty$, $\sigma > 0$. Su gráfica es la figura 7.2.

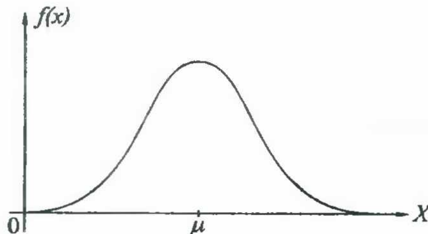


Fig. 7.2 Gráfica de la función de densidad normal

La distribución normal es el modelo probabilístico que se usa más frecuentemente y sirve como una ***buena aproximación*** de muchas distribuciones que tienen aplicaciones importantes.

NOTA. Aceptamos, sin verificar, que

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1.$$

Luego, si se hace $z = (x-\mu)/\sigma$, se tiene $-\infty < z < +\infty$, $dx = \sigma dz$, y

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1.$$

7.2.2.1 Propiedades de la distribución normal

A partir de su gráfica y del análisis de su función de densidad, la curva normal tiene las siguientes propiedades:

1. **Es simétrica** con respecto al eje vertical $X = \mu$, ya que $f(x)$ sólo depende de x mediante la expresión $(x-\mu)^2$, luego,

$$f(\mu + x) = f(\mu - x)$$

2. Tiene ***valor máximo*** en $x = \mu$ (su moda). Este valor máximo es:

$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

3. Tiene al eje X como una ***asíntota horizontal***, ya que

$$\lim_{x\to-\infty} f(x) = 0 \quad \text{y} \quad \lim_{x\to+\infty} f(x) = 0.$$

4. Tiene **puntos de inflexión** en $x = \mu - \sigma$, y $x = \mu + \sigma$, por tanto, es cóncava hacia abajo en el intervalo $\mu - \sigma < x < \mu + \sigma$, y cóncava hacia arriba en cualquier otra parte.

5. El **área total** que encierra la curva y el eje X es igual a 1.

TEOREMA 7.9.

Si la variable aleatoria X tiene distribución normal $N(\mu,\sigma^2)$, entonces,

a) $E(X)=\mu$, b) $Var(X)=\sigma^2$.

PRUEBA.

a) $E(X)=\dfrac{1}{\sigma\sqrt{2\pi}}\displaystyle\int_{-\infty}^{+\infty} xe^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$

Haciendo, $z=(x-\mu)/\sigma$, resulta $-\infty<z<+\infty$, y $dx=\sigma dz$, luego,

$$E(X)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(\sigma z+\mu)e^{-z^2/2}dz$$

$$E(X)=\frac{1}{\sqrt{2\pi}}\sigma\int_{-\infty}^{+\infty} ze^{-z^2/2}dz+\mu\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz$$

La primera integral es igual a cero, por integración directa o por el hecho de que el integrando $h(z)=ze^{-z^2/2}$ es una función impar, esto es, satisface la relación $h(z)=-h(-z)$ y en consecuencia su integral es igual a cero. La segunda integral es igual a 1. Luego, $E(X)=\mu$.

b) $Var(X)=E\left[(X-\mu)^2\right]=\dfrac{1}{\sigma\sqrt{2\pi}}\displaystyle\int_{-\infty}^{+\infty}(x-\mu)^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$

Haciendo, $z=(x-\mu)/\sigma$, resulta $-\infty<z<+\infty$, y $dx=\sigma dz$, luego,

$$E\left[(X-\mu)^2\right]=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} z^2e^{-z^2/2}dz$$

Además si $u=z$ y $dv=ze^{-z^2/2}$, se obtiene $v=-e^{-z^2/2}$, y

$$E\left[(X-\mu)^2\right]=\frac{\sigma^2}{\sqrt{2\pi}}\left[-ze^{-z^2/2}\right]_{-\infty}^{+\infty}+\sigma^2\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz=0+\sigma^2(1)$$

Luego, $Var(X)=\sigma^2$.

7.2.2.2 Distribución normal estándar y uso de la tabla normal

Si la variable aleatoria X tiene distribución normal $N(\mu,\sigma^2)$, entonces, la **variable aleatoria estándar** $Z=(X-\mu)/\sigma$, tiene distribución normal $N(0,1)$.

En efecto, la variable estándar Z tiene media igual a cero y varianza igual a uno, esto es,

$$E(Z)=0 \quad \text{y} \quad Var(Z)=1$$

Además, la probabilidad:

$$P[X \leq x_1]=\int_{-\infty}^{x_1}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx\,.$$

Si X toma el valor x, entonces, el correspondiente valor de Z es $z=(x-\mu)/\sigma$. Si X toma el valor x_1, entonces, el correspondiente valor de Z es $z_1=(x_1-\mu)/\sigma$. También $-\infty<z<+\infty$, y $dx=\sigma dz$, luego,

$$P[X \leq x_1]=\int_{-\infty}^{x_1}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx=\int_{-\infty}^{z_1}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz=P[Z\leq z_1].$$

La función de densidad y la función de distribución acumulada de la normal estándar son respectivamente:

$$\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-z^2/2}, \quad \text{y,} \quad \Phi(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$$

La gráfica de $\Phi(z)$ es la parte sombreada de la figura 7.3.

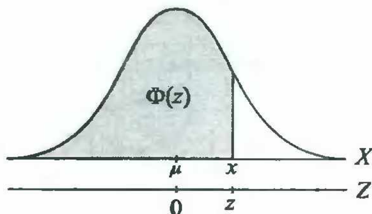


Fig. 7.3 Distribución normal estándar

Tabulación de la distribución normal

La probabilidad:

$$P[a \le Z \le b] = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

no puede calcularse por métodos de integración directa, puesto que no se puede encontrar una función cuya derivada sea igual a $e^{-z^2/2}$. Pero, usando métodos de integración numérica se han *tabulado* los valores de la *función de distribución acumulada de la normal estándar*:

$$\Phi(z) = P[Z \le z] = \int_{-\infty}^{z} \phi(t)dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Las áreas $\Phi(z)$ que se representan en la parte sombreada de la figura 7.3, tienen las siguientes propiedades:

1) $P[a \le Z \le b] = \Phi(b) - \Phi(a)$.

2) $\Phi(-z) = 1 - \Phi(z)$, esto es, es simétrica respecto a la vertical $Z = 0$.

3) $P[-a \le Z \le a] = \Phi(a) - \Phi(-a) = \Phi(a) - (1 - \Phi(a)) = 2\Phi(a) - 1$.

4) Si la variable aleatoria X tiene distribución $N(\mu, \sigma^2)$, entonces, la variable aleatoria $Z = (X - \mu)/\sigma$ tiene distribución $N(0,1)$, luego,

$$P[a \le X \le b] = P\left[\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

En la tabla de probabilidades normal estándar se puede encontrar la probabilidad $1-\alpha$, o el valor $z_{1-\alpha}$ de Z mediante la relación:

$$P[Z \le z_{1-\alpha}] = 1 - \alpha \quad \text{o} \quad \Phi(z_{1-\alpha}) = 1 - \alpha$$

como se indica en la figura 7.4.

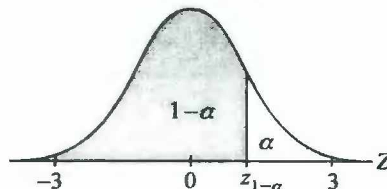


Fig. 7.4 Areas de la distribución normal estándar

EJEMPLO 7.14.

Utilizando la tabla de probabilidades normal hallar

a) $P[Z \leq 1.2]$, b) $P[0.81 \leq Z \leq 1.94]$, c) $P[Z \leq -1.28]$,

d) $P[-0.46 \leq Z \leq 2.21]$, e) $P[Z \geq -0.68]$,

f) $P[-2.04 \leq Z \leq -1.98]$, g) $P[Z \leq 1.676]$.

SOLUCION.

a) $\Phi(1.20) = 0.8849$.

b) $\Phi(1.94) - \Phi(0.81) = 0.9738 - 0.7910 = 0.1828$.

c) $\Phi(-1.28) = 1 - \Phi(1.28) = 1 - 0.8997 = 0.1003$.

d) $\Phi(2.21) - \Phi(-0.46) = \Phi(2.21) + \Phi(0.46) - 1 = 0.9864 + 0.6772 - 1 = 0.6636$.

e) $P[Z \geq -0.68] = P[Z \leq 0.68] = \Phi(0.68) = 0.7517$.

f) $\Phi(2.04) - \Phi(1.98) = 0.9793 - 0.9761 = 0.0032$.

g) Un procedimiento es aproximar z a dos decimales, en este caso,

$$P[Z \leq 1.676] \cong P[Z \leq 1.68] = 0.9535.$$

Si se requiere de mayor precisión, se puede usar el siguiente proceso de **aproximación por interpolación lineal** (ya que se interpola en el tercer decimal y por tanto el pedazo de curva es suavizada por un segmento de longitud muy pequeña):

valores de Z: $1.67 \leq 1.676 \leq 1.68$

áreas respectivas: $0.9525 \leq \Phi(1.676) \leq 0.9535$

Luego, $\dfrac{\Phi(1.676) - 0.9525}{0.9535 - 0.9525} = \dfrac{1.676 - 1.67}{1.68 - 1.67}$, resultando $\Phi(1.676) = 0.9531$.

EJEMPLO 7.15.

Utilizando la tabla de probabilidades normal, hallar el valor de z tal que:

a) $P[Z \leq z] = 0.8621$, b) $P[Z \leq z] = 0.2236$,

c) $P[-z \leq Z \leq z] = 0.9500$ d) $P[-z \leq Z \leq z] = 0.9900$,

e) $P[-z \leq Z \leq z] = 0.9000$.

SOLUCION.

a) $z = 1.09$

b) $z = -0.76$
c) $z = \pm 1.96$
d) $P[-z \le Z \le z] = 2\Phi(z) - 1 = 0.9900$, entonces, $\Phi(z) = 0.9950$

Por interpolación se tiene:

valores de Z: $2.57 \le z \le 2.58$
áreas respectivas: $0.9949 \le 0.9950 \le 0.9951$

Entonces, $\dfrac{z - 2.57}{2.58 - 2.57} = \dfrac{0.9950 - 0.9949}{0.9951 - 0.9949}$, resultando $z = 2.575$.

e) Análogamente $z = 1.645$.

EJEMPLO 7.16.

Si X tiene distribución $N(\mu, \sigma^2)$,

a) Verificar que el área comprendida

a1) Entre $\mu - \sigma$ y $\mu + \sigma$ es 68.26%.
a2) Entre $\mu - 2\sigma$ y $\mu + 2\sigma$ es 95.44%.
a3) Entre $\mu - 3\sigma$ y $\mu + 3\sigma$ es 99.74%.

b) Calcular el valor de k tal que:

$$P[\mu - k\sigma \le X \le \mu + k\sigma] = 0.5000.$$

SOLUCION.

a) $P[\mu - k\sigma \le X \le \mu + k\sigma] = P\left[-k \le \dfrac{X - \mu}{\sigma} \le k\right] = P[-k \le Z \le k] = 2\Phi(k) - 1.$

Luego, $P[\mu - k\sigma \le X \le \mu + k\sigma] = 2\Phi(k) - 1$.

En particular, se tiene:

a1) $P[\mu - \sigma \le X \le \mu + \sigma] = 2\Phi(1) - 1 = 2(0.8413) - 1 = 0.6826$.

a2) $P[\mu - 2\sigma \le X \le \mu + 2\sigma] = 2\Phi(2) - 1 = 2(0.9772) - 1 = 0.9544$.

a3) $P[\mu - 3\sigma \le X \le \mu + 3\sigma] = 2\Phi(3) - 1 = 2(0.9987) - 1 = 0.9974$.

NOTA. Si la distribución de X es no normal, por Chebyshev se tiene

$$P[\mu - k\sigma \le X \le \mu + k\sigma] \ge 1 - \frac{1}{k^2}, \quad k > 0$$

b) De $P[\mu - k\sigma \leq X \leq \mu + k\sigma] = 2\Phi(k) - 1 = 0.5000$, resulta $\Phi(k) = 0.7500$.

Por otra parte, por interpolación, se tiene:

valores de Z : $0.67 \leq k \leq 0.68$

áreas respectivas : $0.7486 \leq 0.7500 \leq 0.7517$.

Entonces, $$\frac{k - 0.67}{0.68 - 0.67} = \frac{0.7500 - 0.7486}{0.7517 - 0.7486}.$$

de donde resulta $k = 0.67 + 0.0045 = 0.6745$.

EJEMPLO 7.17.

Suponga que el ingreso familiar mensual en una comunidad tiene distribución normal con media \$600 y desviación estándar \$100.

a) Calcular la probabilidad de que el ingreso de una familia escogida al azar sea menor que \$400.

b) Si el 5% de las familias con mayores ingresos deben pagar un impuesto, ¿a partir de que ingreso familiar se debe pagar el impuesto?.

SOLUCION.

Sea X la variable que representa los ingresos familiares mensuales . La distribución de X es $N(600, (100)^2)$.

a) $P[X < 400] = P[Z < \frac{400 - 600}{100}] = P[Z < -2] = 0.0228$

b) Se debe hallar K tal que, $P[X \geq K] = 0.05$ o $P[X < K] = 0.95$, entonces,

$0.95 = P[X < K] = P[Z < \frac{K - 600}{100}]$, de donde resulta $\frac{K - 600}{100} = 1.645$,

$K = 764.5$.

EJEMPLO 7.18.

Suponga que la duración X de los focos que produce una compañía se distribuye normalmente. Si el 18.41% de estos focos duran menos de 8.2 meses y el 6.68% duran al menos 13 meses.

a) Calcular la media y la varianza de la duración de los focos.

b) Hallar el cuartil Q_1 de la distribución.

c) Si el costo de cada foco es de S/.10 y se vende en S/.25, pero garantizando la devolución total del dinero si dura menos de 8 meses, ¿cuál es la utilidad esperada por foco?.

d) Si mediante un nuevo proceso se aumenta la duración en 10% más 15 días, ¿qué porcentaje de focos duran ahora menos de 7 meses?.

SOLUCION.

Sea la variable aleatoria *X*, la duración de los focos. La distribución de X es normal con media μ y varianza σ^2.

a) Si el 18.41% duran menos de 8.2 meses, entonces $P[X < 8.2] = 0.1841$, luego,

$$0.1841 = P[X < 8.2] = P\left[Z < \frac{8.2-\mu}{\sigma}\right] = \Phi\left(\frac{8.2-\mu}{\sigma}\right)$$

de donde se obtiene $\dfrac{8.2-\mu}{\sigma} = -0.9$, y $\mu - 0.9\sigma = 8.2$.

Si el 6.68% duran al menos 13 meses, entonces, $P[X \geq 13] = 0.0668$,

luego, $$0.9332 = P[X < 13] = P\left[Z < \frac{13-\mu}{\sigma}\right] = \Phi\left(\frac{13-\mu}{\sigma}\right)$$

de donde resulta $\dfrac{13-\mu}{\sigma} = 1.50$, y $\mu + 1.5\sigma = 13$.

Resolviendo simultáneamente las ecuaciones resultantes,

$$\mu + 1.5\sigma = 13$$
$$\mu - 0.9\sigma = 8.2$$

se obtienen: $\mu = 10$, y $\sigma = 2$.

b) El cuartil 1, Q_1, satisface la condición $P[X \leq Q_1] = 0.25$, entonces,

De $P[Z < (Q_1 - 10)/2] = 0.25$, resulta $\dfrac{Q_1 - 10}{2} = -0.67$, $Q_1 = 8.66$.

c) La utilidad $U = -10$ si $X < 8$, $U = 15$ si $X \geq 8$. La utilidad esperada es:

$$E(U) = -10P[X < 8] + 15P[X \geq 8] = -10P[Z < -1] + 15P[Z \geq -1]$$

$$E(U) = -10 \times 0.1587 + 15 \times 0.8413 = 11.03$$

d) Sea *Y* la duración con el nuevo proceso, entonces $Y = 1.1X + 0.5$, y

$$P[Y < 7] = P[X < 5.91] = P[Z < -2.05] = 0.0202$$

EJEMPLO 7.19.

En un proceso de fabricación de arandelas se estima que el diámetro interior varía de 12.319 mm. a 13.081 mm. El propósito para el que se destinan estas arandelas permiten una tolerancia en el diámetro de 12.6 mm. a 12.9 mm., de otro

modo las arandelas se consideran defectuosas. Si el diámetro de las arandelas es una variable aleatoria que se distribuye normalmente con media 12.7 mm.

a) Calcular el porcentaje de arandelas defectuosas producidas
b) y si se producen 2000 arandelas, calcular el número esperado de arandelas aceptables.

SOLUCION.

Sea la variable aleatoria X el diámetro interior de las arandelas. Se sabe que X tiene distribución $N(12.7,\sigma^2)$.

Para determinar la desviación estándar de la distribución observemos que casi el 100% del área total de la distribución se incluye en el intervalo $\mu \mp 3\sigma$, ya que

$$P[\mu-3\sigma \leq X \leq \mu+3\sigma]=0.9974.$$

Por lo tanto, el rango R de valores de la variable X es aproximadamente:

$$R \cong (\mu+3\sigma)-(\mu-3\sigma): \ R \cong 6\sigma \ \text{ o } \ \sigma \cong R/6.$$

Por otro lado, de los datos, $R=X_{max}-X_{min}=13.081-12.319 \cong 0.762$.

Luego, $\sigma \cong R/6 = 0.762/6 = 0.127$ mm.

La probabilidad de arandelas no defectuosas es

$$P[12.6 \leq X \leq 12.9]=P\left[\frac{12.6-12.7}{0.127} \leq Z \leq \frac{12.9-12.7}{0.127}\right]$$

$$=P[-0.79 \leq Z \leq 1.57]=\Phi(1.57)-\Phi(-0.79)$$

$$=0.9418+0.7852-1=0.7270.$$

a) El porcentaje de arandelas defectuosas es 100% − 72.7% = 27.3%.

b) El número esperado de arandelas aceptables es 2000×0.7270 = 1454.

EJEMPLO 7.19 B

Sea X una variable aleatoria continua cuyo modelo de probabilidad es:

$$f(x)=\frac{1}{\sqrt{\pi/2}}e^{-2(X^2-X+0.25)}$$

Calcule la probabilidad de que X esté en el intervalo [0 , 1.5]

SOLUCION.

La función de densidad de probabilidad corresponde a: $X \sim N(0.5, 0.5^2)$.

Entonces, $P[0 \leq X \leq 1,.5]=P[-1 \leq Z \leq 2]=0.8185$

NOTA (Propiedad reproductiva de la normal).

Si $X_1, X_2, ... X_k$, son k variables aleatorias independientes, tales que $X_i \sim N(\mu_i, \sigma_i^2)$, para cada $i = 1,2,...,k$, entonces, la variable aleatoria $c_1 X_1 + c_2 X_2 + ... + c_k X_k$ (donde $c_1, c_2, ..., c_k$ son constantes reales) está distribuida normalmente

con media $c_1\mu_1 + c_2\mu_2 + ... + c_k\mu_k$ y con varianza $c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + ... + c_k^2\sigma_k^2$.

En particular, si las variables aleatorias X, Y tienen las distribuciones normales respectivas: $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, entonces,

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \quad \text{y} \quad X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

EJEMPLO 7.20.

Un producto cuyo peso se distribuye normalmente con un promedio de 250 gramos y una desviación estándar de 10 gramos es embalado en cajas de 4 docenas cada una. Se supone que el peso de las cajas vacías se distribuyen normalmente con un promedio de 350 gramos y una desviación estándar de 15 gramos. Si los pesos del producto y de la caja son independientes,

a) Calcular la probabilidad de que una caja llena pese menos de 12.5 kg.
b) Cuántas cajas llenas serán necesarias para que el peso total sea al menos 110,724.56 gramos con probabilidad de 0.9772.

SOLUCION.

Sean $X_1, X_2, ..., X_{48}$ las variables aleatorias que representa el peso de los 48 productos, e Y la variable aleatoria que representa el peso de la caja. Se sabe que cada $X_i \sim N(250, 10^2)$ y que $Y \sim N(350, 15^2)$. Entonces, el peso de la caja llena, es la variable aleatoria:

$$W = \sum_{i=1}^{48} X_i + Y$$

cuya distribución es normal con media:

$$E(W) = \sum_{i=1}^{48} E(X_i) + E(Y) = 48(250) + 350 = 12{,}350 \text{ gramos.}$$

y con varianza

$$Var(W) = \sum_{i=1}^{48} Var(X_i) + Var(Y) = 48(100) + 225 = 5025 \text{ gramos}^2.$$

Luego,

a) $P[W < 12{,}500] = P\left[\dfrac{W - 12{,}350}{70.89} > \dfrac{12{,}500 - 12{,}350}{70.89}\right] = P[Z < 2.12] = 0.9830.$

b) Sea $U_n = \sum_{i=1}^{n} W_i$, el peso de n cajas llenas, entonces, la variable aleatoria U_n normal con media $12{,}350n$ y con varianza $5025n$.

Se debe calcular n tal que $P[U_n \geq 10{,}724.56] = 0.9772$, luego,

$$0.0228 = P[U_n < 110{,}724.56] = P\left[Z < \frac{110{,}724.56 - 12{,}350n}{\sqrt{5025n}}\right]$$

de donde resulta, $\dfrac{110{,}724.56 - 12{,}350n}{\sqrt{5025n}} = -2.00$, cuya solución es $n = 9$.

7.2.3 Distribuciones gamma, exponencial y chi-cuadrado

7.2.3.1 Función Gamma.

Definición. La **función gamma** denotada por, Γ , se define por:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx,$$

donde α es un número real positivo.

Propiedades.

1) Si $\alpha > 1$, entonces, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

En efecto, haciendo $u = x^{\alpha - 1}$, $dv = e^{-x}dx$, en la función gamma, e integrando por partes, se obtiene,

$$\Gamma(\alpha) = \lim_{b \to +\infty} \left[-x^{\alpha-1} e^{-x} \right]_0^b + (\alpha - 1) \int_0^{\infty} x^{\alpha-2} e^{-x} dx = 0 + (\alpha - 1) \int_0^{\infty} x^{\alpha-2} e^{-x} dx$$

Luego, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

2) $\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$.

3) $\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} x^{-1/2} e^{-x} = \sqrt{\pi}$.

4) Si α es igual a un número entero: $n \geq 1$, entonces, $\Gamma(n) = (n-1)!$

En efecto, aplicando: $\Gamma(n) = (n-1)\Gamma(n-1)$, repetidamente,
resulta,

$$\Gamma(n) = (n-1)(n-2)(n-3)...\Gamma(1).$$

Luego, $\Gamma(n) = (n-1)!$.

7.2.3.2 Distribución gamma

Definición. Se dice que la variable aleatoria continua X tiene **distribución gamma,** con parámetros α, β y se representa por $X \sim \Gamma(\alpha, \beta)$, si su función de densidad es:

$$f(x) = \begin{cases} \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & \text{si} \quad x \geq 0 \\ 0, & \text{si} \quad x < 0 \end{cases}$$

donde α y β son constantes positivas.

La figura 7.5 es la gráfica de la distribución gamma cuando $\beta = 1$, para $\alpha = 1$, $\alpha = 2$, $\alpha = 4$.

TEOREMA 7.10.

Si $X \sim \Gamma(\alpha, \beta)$, entonces, a) Media: $\mu = \dfrac{\alpha}{\beta}$, b) Varianza: $\sigma^2 = \dfrac{\alpha}{\beta^2}$.

PRUEBA.

a) $E(X) = \int_0^{\infty} x \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \int_0^{\infty} \frac{(\beta x)^{\alpha}}{\Gamma(\alpha)} e^{-\beta x} dx$

Si $\beta x = \mu$, entonces, $0 < u < +\infty$, y $dx = du/\beta$, luego,

$$E(X) = \frac{1}{\beta\Gamma(\alpha)} \int_0^{\infty} u^{\alpha} e^{-u} du = \frac{1}{\beta} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} = \frac{\alpha}{\beta}.$$

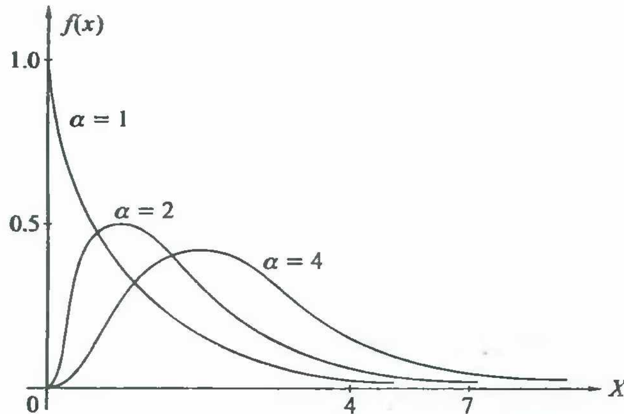


Fig. 7.5. Gráfica de la distribución gamma con $\beta = 1$.

b) $E(X^2) = \int_0^{\infty} x^2 \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \frac{1}{\beta} \int_0^{\infty} \frac{(\beta x)^{\alpha+1}}{\Gamma(\alpha)} e^{-\beta x} dx$

Si $\beta x = u$ entonces, $dx = du/\beta$ y

$$E(X^2) = \frac{1}{\beta^2\Gamma(\alpha)} \int_0^{\infty} u^{\alpha+1} e^{-u} du = \frac{1}{\beta^2} \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \frac{\alpha(\alpha+1)\Gamma(\alpha)}{\beta^2\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\beta^2}$$

Luego,

$$Var(X) = E\left(X^2\right) - \left(E(X)\right)^2 = \frac{\alpha(\alpha+1)}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}.$$

7.2.3.3 Distribución exponencial

Definición. Se dice que la variable aleatoria continua X tiene **distribución exponencial con parámetro** β ($\beta > 0$), y se describe $X \sim Exp(\beta)$, si su función de densidad de probabilidades es:

$$f(x) = \begin{cases} \beta e^{-\beta x}, & \text{si } x \geq 0 \\ 0, & \text{si } x < 0 \end{cases}$$

Su gráfica es la figura 7.6. Es un modelo apropiado a vida útil de objetos.

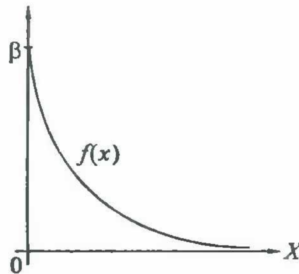


Fig. 7.6. Gráfica de la distribución exponencial

NOTA. La distribución exponencial es un caso particular de la distribución gamma cuando $\alpha = 1$. Luego, si la variable aleatoria X tiene distribución exponencial con parámetro β o $X \sim \Gamma(1, \beta)$, entonces tiene,

$$\text{a) Media: } \mu = \frac{1}{\beta}, \quad \text{b) Varianza: } \sigma^2 = \frac{1}{\beta^2}$$

Por otra parte, su función de distribución acumulada en el intervalo $[0, +\infty[$ es: ,

$$F(x) = P[X \leq x] = \int_0^x \beta e^{-\beta x} dt = 1 - e^{-\beta x}, \quad 0 \leq x < \infty.$$

Observar también que: $P[X > x] = 1 - P[X \leq x] = e^{-\beta x}, \quad 0 \leq x < \infty.$

Además, para números reales positivos s y t cualesquiera, se verifica:

$$P[X > s + t / X > s] = P[X > t]$$

En efecto,

$$P[X > s + t / X > s] = \frac{P[X > s + t]}{P[X > s]} = \frac{e^{-\beta(s+t)}}{e^{-\beta s}} = e^{-\beta t} = P[X > t]$$

EJEMPLO 7.21.

El tiempo durante el cual cierta marca de batería trabaja en forma efectiva hasta que falle (tiempo de falla) se distribuye según el modelo exponencial con un tiempo promedio de fallas igual a 360 días.

a) ¿Qué probabilidad hay que el tiempo de falla sea mayor que 400 días?. Esta probabilidad es conocida también como **confiabilidad** de la batería.

b) Si una de estas baterías ha trabajado ya 400 días, ¿qué probabilidad hay que trabaje más de 200 días más?.

c) Si se están usando 5 de tales baterías calcular la probabilidad de que más de dos de ellas continúen trabajando después de 360 días.

SOLUCION.

Sea X = el tiempo que trabaja la batería hasta que falle. El tiempo promedio de falla es de 360 días. Entonces, $X \sim Exp(\beta = 1/360)$ y su función de densidad es:

$$f(x) = \frac{1}{360} e^{-x/360}, \quad 0 \leq x < \infty$$

a) $P[X > 400] = e^{-400/360} = 0.329$

b) Si la batería ya trabajó 400 días, quiere decir que su tiempo de falla es mayor que 400 días. Luego,

$$P[X > 400 + 200 / X > 400] = P[X > 200] = e^{-200/360} = 0.574.$$

c) La probabilidad de que una batería trabaje más de 360 días es:

$$p = P[X > 360] = e^{-360/360} = e^{-1} = 0.368.$$

Sea Y = # de baterías de 5 que siguen trabajando después de 360 días, entonces, $Y \sim B(5, p)$, y

$$P[Y \geq 3] = 1 - P[Y \leq 2] = 1 - \sum_{k=0}^{2} C_k^5 (0.368)^k (0.632)^{5-k} = 0.26376$$

NOTA. (Modelo exponencial y modelo de Poisson).

Si X es el número de veces que ocurre un evento en un período dado con promedio λ de ocurrencias, (esto es si X tiene distribución de Poisson con promedio λ), entonces, el tiempo que transcurre hasta que ocurre el primer evento de Poisson es una variable aleatoria continua que tiene distribución exponencial con parámetro $\beta = \lambda$.

EJEMPLO 7.22.

a) Sea X el numero de eventos de Poisson que ocurren en un intervalo de tiempo $[0, t]$ con promedio λt (λ es el promedio en el intervalo [0, 1]). Si T es el tiempo que transcurre hasta que ocurra el primer evento de Poisson verificar que T tiene distribución exponencial con parámetro $\beta = \lambda$.

b) Suponga que el número de llamadas que llegan a una central telefónica tiene distribución de Poisson con un promedio de 5 llamadas por minuto.¿Cuál es la probabilidad de que la primera llamada ocurra en a lo más 30 segundos de haber sido puesto en marca de la central?.

SOLUCION.

a) X = # eventos de Poisson que ocurren en el intervalo [0, t]. Entonces,

$$P[X = k] = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \quad k{=}0,1,2,\text{etc..}$$

Sea T=tiempo que transcurre hasta que ocurre el primer evento de Poisson. El rango de T, es el intervalo $R_T = [0,+\infty[$.

Primero se halla la función de distribución acumulada $F(t)$ de T y luego derivando ésta se halla la función de densidad. Para $t \in R_T$ se tiene :

$$F(t) = P[T \le t] = 1 - P[T > t] = 1 - P[X = 0] = 1 - e^{-\lambda t}.$$

Donde, el evento $[T > t]$ indica que el primer evento de Poisson ocurre después de t (o en el intervalo $]t,+\infty[$) lo que es lo mismo que no ocurre ningún evento de Poisson en el intervalo [0, t]. Es decir, $[T > t] = [X = 0]$. Finalmente,

$$f(t) = \frac{dF(t)}{dt} = \lambda e^{-\lambda t}, \quad t \ge 0.$$

b) Sea X: # de llamadas que ocurren con un promedio de $\lambda = 5$ por minuto.
Sea T el tiempo en minutos que transcurre hasta que ocurra la primera llamada.
Entonces, T es exponencial con parámetro $\beta = \lambda = 5$. Su función de densidad es:

$$f(t) = 5e^{-5t}, \quad 0 \le t < +\infty$$

La probabilidad que la primera llamada ocurra en el intervalo $[0, 1/2]$ es:

$$P[T \le 1/2] = \int_0^{1/2} 5e^{-5t}dt = 1 - e^{-5/2} = 0.918.$$

Observar que: $P[T \le 1/2] = 1 - P[T > 1/2] = 1 - P[X = 0] = 1 - e^{-5\times(1/2)}$

El evento; $[T > 1/2]$: "ocurre la primera llamada después de 1/2 minuto" es igual al evento $[X = 0]$ "no ocurre ninguna llamada en el periodo $[0, 1/2]$ minutos con promedio $\lambda t = 5 \times (1/2)$ "

NOTA. (Modelo Gamma y modelo de Poisson)
La distribución gamma describe la función de densidad de la variable aleatoria que representa el tiempo que transcurre hasta que ocurra un número específico de eventos de Poisson con promedio λ. Este número específico es el parámetro α, y $\beta = \lambda$ en la función de densidad gamma.

EJEMPLO 7.23.
El número de automóviles que llegan a un surtidor de gasolina se distribuyen según el modelo de Poisson con un promedio de 2 autos por minuto. ¿Qué probabilidad hay de que transcurran hasta dos minutos hasta que llegue el segundo auto, después de haber puesto en marcha el surtidor?.
SOLUCION.
Sea X = # de autos que llegan al surtidor con promedio $\lambda = 2$ en el período de un minuto. X se distribuye según la ley de Poisson.
Sea T el tiempo que transcurre hasta que llegue el segundo auto, después de haber sido puesto en marcha el surtidor. Entonces, T tiene distribución gamma con parámetros $\alpha = 2$ y $\beta = 2$. Esto es, $T \sim \Gamma(\alpha = 2, \beta = 2)$. La probabilidad de que transcurran hasta 2 minutos hasta que llegue el segundo auto es:

$$P[T \leq 2] = 4\int_0^2 te^{-2t}dt = 1 - e^{-4}(1+4) = 0.908.$$

Observe que: $P[T \leq 2] = 1 - P[T > 2] = 1 - P[0 \leq X \leq 1] = 1 - e^{-4}(1+4)$.
El evento; $[T > 2]$: "transcurren más de dos minutos hasta que llegue el segundo auto" es igual al evento $[X \leq 1]$:"llega a lo más un auto en el periodo $[0, 2]$ minutos con promedio $\lambda t = 2 \times 2 = 4$ "

7.2.3.4 Distribución Chi-cuadrado

Definición. Se dice que la variable aleatoria continua X tiene distribución **chi-cuadrado con r grados de libertad**, y se representa por $X \sim \chi^2(r)$, si su función de densidad es:

$$f(x) = \begin{cases} \dfrac{2^{-r/2}}{\Gamma(r/2)} x^{r/2-1} e^{-x/2}, & \text{si } x \geq 0 \\ 0, & \text{si } x < 0 \end{cases}$$

donde r es un número entero positivo.

Su gráfica se da en la figura 7.7 .

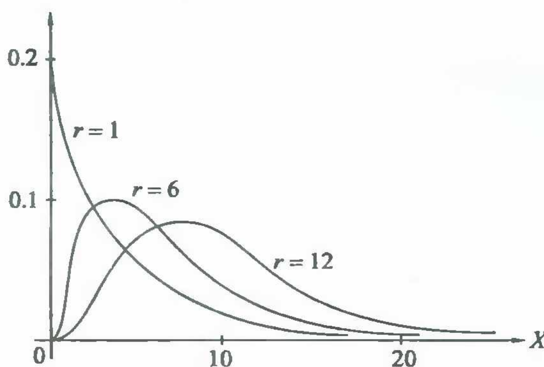


Fig. 7.7. Gráfica de la distribución chi-cuadrado

NOTA. La distribución chi-cuadrado es un caso particular de la distribución gamma, cuando $\alpha = r/2$ y $\beta = 1/2$. Luego, si $X \sim \chi^2(r)$ o si $X \sim \Gamma(r/2, 1/2)$, entonces, su media y su varianza respectivamente son:

a) $\mu = r$ y b) $\sigma^2 = 2r$

PROPIEDADES.

Las siguientes proposiciones (cuya prueba omitimos) son de importancia para el desarrollo de este texto.

1) Si $Z \sim N(0,1)$, entonces, $Z^2 \sim \chi^2(1)$.

2) Si $Z_1, Z_2, ..., Z_r$, son r variables aleatorias independientes tales que $Z_i \sim N(0,1)$, para cada $i = 1,2,3,...,r$, entonces,

$$\sum_{i=1}^{r} Z_i^2 \sim \chi^2(r)$$

Con frecuencia, $\chi^2(r)$, se define como la suma de los cuadrados de r variables aleatorias independientes distribuidas cada una como $N(0,1)$.

3) Propiedad reproductiva de la distribución chi-cuadrado

Si $X_1, X_2, ..., X_k$, son k variables aleatorias independientes tales que $X_i \sim \chi^2(r_i)$, para cada $i = 1,2,3,...,k$, entonces,

$$\sum_{i=1}^{k} X_i^2 \sim \chi^2(r_1 + r_2 + ... + r_k).$$

Uso de la tabla chi-cuadrado

Si la variable aleatoria *X* se distribuye como una chi-cuadrado con *r* grados de libertad, esto es, si $X \sim \chi^2(r)$, entonces, en la tabla de probabilidades chi-cuadrado se puede encontrar una *probabilidad* $1-\alpha$ o un *valor* $c = \chi^2_{1-\alpha,r}$, mediante la relación

$$P[X \leq \chi^2_{1-\alpha,r}] = 1 - \alpha$$

como se indica en la figura 7.8.

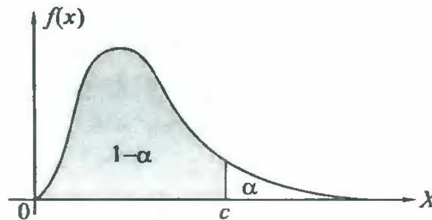


Fig. 7.8. Gráfica chi-cuadrado

EJEMPLO 7.24.

Si $X \sim \chi^2(26)$, determinar:

a) $P[X \leq 17.29]$, b) $P[X \geq 38.89]$,

c) $P[13.84 \leq X \leq 45.64]$, d) $P[X \leq 40]$.

SOLUCION.

a) $P[X \leq 17.29] = 0.10$,

b) $P[X \geq 38.89] = 1 - 0.95 = 0.05$,

c) $P[13.84 \leq X \leq 45.64] = 0.99 - 0.025 = 0.965$.

d) Por interpolación:

Valor de chi-cuadrado:	$38.89 \leq 40 \leq 41.92$
Area acumulada correspondiente:	$0.95 \leq p_0 \leq 0.975$

Luego, $\dfrac{p_0-0.95}{0.975-0.95}=\dfrac{40-38.89}{41.92-38.89}$, de donde resulta $p_0=0.9592$.

EJEMPLO 7.25.

Si $X\sim\chi(r)$, hallar:

a) a tal que $P[X\le a]=0.995$, si $r=30$.

b) a y b tales que $P[a\le X\le b]=0.95$, $P[X>b]=0.025$. si $r=13$

c) a tal que $P[X\le a]=0.015$, si $r=8$.

SOLUCION.

a) $a=53.67$.

b) De $P[X>b]=0.025$, se tiene $P[X\le b]=0.975$, resultando $b=24.74$

Por otra parte, $0.95=P[a\le X\le b]=P[X\le b]-P[X\le a]$,
de donde resulta,

$$P[X\le a]=0.975-0.95=0.025\text{, entonces , } a=5.01.$$

c) Por interpolación:

Area acumulada: $0.010\le 0.015\le 0.025$

valor de chi-cuadrado: $1.65\le a\le 2.18$

luego, $\dfrac{a-1.65}{2.18-1.65}=\dfrac{0.015-0.010}{0.025-0.010}$, de donde se obtiene $a=1.8267$.

7.2.4 Distribución t de Student

Definición. Se dice que una variable aleatoria continua T se distribuye según ***t*-student** (más brevemente según t) **con r grados de libertad** y se representa por $T\sim t(r)$, si su función de densidad es,

$$f(t)=\frac{\Gamma[(r+1)/2]}{\Gamma(r/2)\sqrt{r(\pi)}}\left(1+\frac{t^2}{r}\right)^{-(r+1)/2}\qquad -\infty<t<\infty,$$

donde r es un entero positivo.

Su gráfica es de forma campanoide como se representa en la figura 7.9

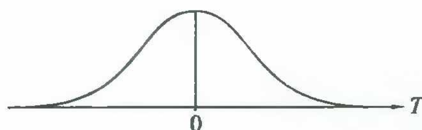


Fig. 7.9. Gráfica de la distribución t-Student

Se demuestra, en particular, que;

"Si Z y V son dos variables aleatorias independientes tales que Z está normalmente distribuida con media cero y varianza 1, y V está distribuida como chi-cuadrado con r grados de libertad, entonces, la variable aleatoria

$$T = \frac{Z}{\sqrt{V/r}}$$

tiene distribución t-Student con r grados de libertad".

Con frecuencia, la distribución t con r grados de libertad es definida como la distribución de la variable T.

La distribución t-Student tiene las siguientes propiedades:

1. Si X tiene distribución t-Student con r grados de libertad, entonces su media y su varianza son respectivamente.

$$\text{a) } \mu = 0, \quad \text{b) } \sigma^2 = \frac{r}{r-2}, \quad r > 2.$$

2. Su gráfica tiene forma de campana de Gauss, simétrica en cero.

3. La varianza de la distribución t es mayor que de la distribución $N(0,1)$. Pero cuando $r \to +\infty$, la varianza de la t tiende a 1.

4. La distribución t *se aproxima* a una distribución $N(0,1)$, cuando $r \to +\infty$. La aproximación es buena, si $r \geq 30$.

Uso de la tabla t-Student

Si la variable aleatoria T tiene distribución t-Student con r grados de libertad, o $T \sim t(r)$, en la tabla de probabilidades t-Student se puede encontrar una *probabilidad* $1-\alpha$ o un valor $c = t_{1-\alpha,r}$ mediante la relación:

$$P[T \leq t_{1-\alpha,r}] = 1-\alpha$$

como se indica en la figura 7.10.

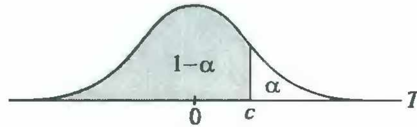


Fig. 7.10. Areas de la distribución t

EJEMPLO 7.26.

Si X tiene distribución t-Student con 18 grados de libertad, hallar:

a) $P[X \leq 2.101]$, b) $P[X > 1.734]$,

c) $P[X \leq -2.878]$, d) $P[-1.330 \leq X \leq 2.552]$, e) $P[X \leq 2]$.

SOLUCION.

a) $P[X \leq 2.101] = 0.975$,

b) $P[X > 1.734] = 1 - 0.95 = 0.05$,

c) $P[X \leq -2.878] = 1 - P[X \leq 2.878] = 1 - 0.995 = 0.005$.

d) $0.990 - (1 - 0.90) = 0.89$.

e) Por interpolación:

Valor t : $1.734 \leq 2 \leq 2.101$

áreas respectivas : $0.95 \leq p_0 \leq 0.975$

Luego, $\dfrac{p_0 - 0.95}{0.975 - 0.95} = \dfrac{2 - 1.734}{2.101 - 1.734}$, de donde resulta, $p_0 = 0.968$.

EJEMPLO 7.27.

Si X tiene distribución t con 10 grados de libertad hallar el valor c tal que:

a) $P[X \leq c] = 0.995$, b) $P[X \leq c] = 0.05$,

c) $P[X > c] = 0.01$ d) $P[-c \leq X \leq c] = 0.95$,

e) $P[X \leq c] = 0.92$.

SOLUCION.

a) $P[X \leq c] = 0.995$, implica, $c = 3.169$.

b) $P[X \leq c] = 0.05$, implica, $c = -1.812$.

c) $P[X > c] = 0.01$, implica, $P[X \leq c] = 0.99$, luego $c = 2.764$.

d) $P[-c \leq X \leq c] = 0.95$, implica $P[X \leq c] = 0.975$, luego $c = 2.228$.
e) Por interpolación:

Valor t : $1.372 \leq c \leq 1.812$

áreas respectivas : $0.90 \leq 0.92 \leq 0.95$

Luego, $\dfrac{c - 1.372}{1.812 - 1.372} = \dfrac{0.92 - 0.90}{0.95 - 0.90}$, de donde resulta $c = 1.548$.

7.2.5 Distribución F

Definición. Se dice que una variable aleatoria continua X se distribuye *según F con r_1 y r_2 grados de libertad* y se representa por $F \sim F(r_1, r_2)$, si su función de densidad es

$$f(x) = \frac{\left(\frac{r_1}{r_2}\right)^{r_1/2} \Gamma\left(\frac{r_1 + r_2}{2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)} \cdot \frac{x^{(r_1/2)-1}}{\left(1 + \frac{r_1 x}{r_2}\right)^{(r_1+r_2)/2}}, \quad 0 \leq x < \infty$$

donde r_1 y r_2, son números enteros positivos. Su gráfica es la figura 7.11.

Se demuestra, en particular, que si U y V son dos variables aleatorias independientes tales que $U \sim \chi^2(r_1)$ y $V \sim \chi^2(r_2)$, entonces, la variable aleatoria:

$$X = \frac{U/r_1}{V/r_2}$$

tiene distribución F con r_1 y r_2 grados de libertad.

Con frecuencia la distribución $F(r_1, r_2)$ se define como la función de densidad de esta variable X.

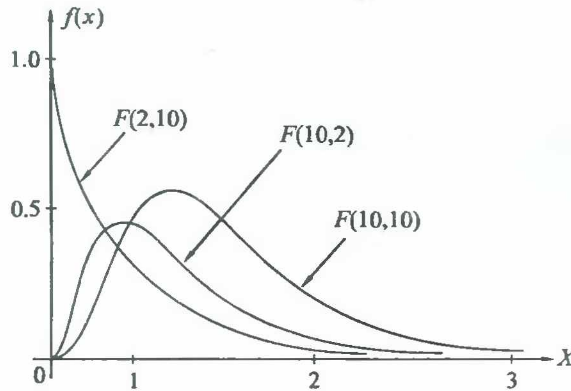


Fig. 7.11. Gráfica de la distribución F

Uso de la tabla F

Si la variable aleatoria $X \sim F(r_1, r_2)$, en la tabla de probabilidades F se puede encontrar una *probabilidad* $1-\alpha$ o un *valor* $c = F_{1-\alpha, r_1, r_2}$, mediante la relación: $P[X \le c] = 1-\alpha$, como se indica en la figura 7.12.

Para determinar valores de F correspondientes a áreas $1-\alpha = 0.005$, 0.01, 0.025, 0.05, o para determinar probabilidades correspondientes a valores de $c < 1$ se usa el teorema siguiente:

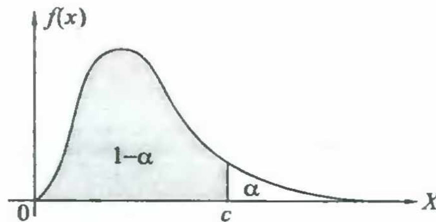


Fig. 7.12. Areas de la distribución F

TEOREMA 7.11.

Si X tiene distribución F con grados de libertad r_1 y r_2, entonces, $1/X$ tiene distribución F con grados de libertad r_2 y r_1, esto es,

$$F_{1-\alpha,r_1,r_2} = \frac{1}{F_{\alpha,r_2,r_1}}$$

PRUEBA.

$$1-\alpha = P\left[F \le F_{1-\alpha,r_1,r_2}\right] = P\left[\frac{U/r_1}{V/r_2} \le c\right] = P\left[\frac{V/r_2}{U/r_1} \ge \frac{1}{c}\right] = 1 - P\left[\frac{V/r_2}{U/r_1} \le \frac{1}{c}\right]$$

Luego, $$P\left[\frac{V/r_2}{U/r_1} \le \frac{1}{c}\right] = \alpha.$$

Por otra parte, $(V/r_2)/(U/r_1)$ se distribuye según $F(r_2, r_1)$, entonces, $\frac{1}{c} = F_{\alpha,r_2,r_1}$, esto es,

$$F_{1-\alpha,r_1,r_2} = \frac{1}{F_{\alpha,r_2,r_1}}$$

EJEMPLO 7.28.

Si $X \sim F(4,5)$ hallar:

a) $P[X \le 7.39]$, b) $P[X > 11.4]$,

c) $P[X \le 8]$, d) $P[X \le 0.0645]$.

SOLUCION.

a) $P[X \le 7.39] = 0.975$,

b) $P[X > 11.4] = 1 - P[X \le 11.4] = 1 - 0.99 = 0.01$.

c) Por interpolación:

valor de F : $7.39 \le 8 \le 11.4$

Area correspondiente : $0.975 \le p_0 \le 0.99$

Luego, $\frac{p_0 - 0.975}{0.99 - 0.975} = \frac{8 - 7.39}{11.4 - 7.39}$, de donde resulta $p_0 = 0.977$.

d) Hay que hallar $1-\alpha$ tal que $F_{1-\alpha,4,5} = 0.0645$. Se sabe que:

$$F_{\alpha,5,4} = \frac{1}{F_{1-\alpha,4,5}} = \frac{1}{0.0645} = 15.5$$

de donde se obtiene $\alpha = 0.99$, luego, $1-\alpha = 0.01$

EJEMPLO 7.29.
Si $X \sim F(6,10)$, buscar el valor de c tal que:

a) $P[X \leq c] = 0.99$, b) $P[X \geq c] = 0.05$,

c) $P[X \leq c] = 0.025$.

SOLUCION.

a) $P[X \leq c] = 0.99$ implica $c = 5.39$.

b) De $P[X \geq c] = 0.05$, se obtiene $P[X \leq c] = 95$, luego, $c = 3.22$.

c) $F_{0.025,6,10} = \dfrac{1}{F_{0.975,10,6}} = \dfrac{1}{5.46} = 0.183.$

EJEMPLO 7.30.
Si $X \sim F(15,12)$, hallar a y b tales que $P[a \leq X \leq b] = 0.90$, sabiendo que $P[X > b] = 0.05$.

SOLUCION.

De $P[X > b] = 0.05$, se obtiene $P[X \leq b] = 0.95$, entonces, $b = 2.62$.
De $0.90 = P[a \leq X \leq b] = P[X \leq b] - P[X \leq a]$, resulta $P[X \leq a] = 0.05$. Luego,

$$a = F_{0.05,15,12} = \frac{1}{F_{0.95,12,15}} = \frac{1}{2.48} = 0.4.$$

EJERCICIOS

Uniforme

1. Suponga que el tiempo X, en minutos, que demora una tarea tiene distribución uniforme en $[1,5]$. Si el costo C para terminar la tarea es función del tiempo y es dada por la expresión

$$C = 10 + X + 3X^2$$

calcular el valor esperado del costo.

Rp. $E(X)=3$, $V(X)=4/3$, $E(C)=10+3E(X)+E(X^2)=10+3+3(V(X)+\mu^2)]=44$.

2. Los ómnibus llegan a un paradero especifico en intervalos de 15 minutos comenzando a las 7 a.m.. Esto es, los ómnibus llegan al paradero a las 7, 7.15, 7.30, 7.45, y así sucesivamente. Si un pasajero llega al paradero en un tiempo que está uniformemente distribuido entre las 7 a.m. y 8 a.m., calcular la probabilidad de que el pasajero espere:

a) A lo más 5 minutos por un ómnibus

b) más de 10 minutos por un ómnibus

Rp. a) $P[10\leq T\leq 15$ o $25\leq T\leq 30$ o $40\leq T\leq 45$ o $55\leq T\leq 60]=4(5/60)=1/3$. b) $4(5/60)=1/3$.

3. Un vendedor tiene un sueldo fijo de S/.400 más una comisión del 5% sobre el importe de las ventas que realiza. Si el importe de las ventas tiene una distribución uniforme entre 0 y 3400 nuevos soles,

a) Hallar el ingreso medio del vendedor, ¿con qué probabilidad obtendría al menos ese monto?.

b) Se le ofrece como ingreso único el 25% de sus ventas, si como mínimo quiere ganar $480, ¿le conviene la propuesta?.

Rp. $Y=400+0.05X$, $X\sim U[0,3400]$, a) $E(Y)=485$, Prob=0.5 b) $P[Y>480]>P[0.25X>480]$. No.

4. Una de las clases de un determinado profesor está programado para comenzar a las 8.10 a.m., pero él comienza su clase en un tiempo T que se distribuye uniformemente en el intervalo de 8.05 a.m. a 8.15 a.m. ¿Cuál es la probabilidad de que él comience su clase,

a) al menos dos minutos más temprano?.

b) a lo más tres minutos más tarde?.

Rp. a) 3/10, b) 3/10.

5. La llegada de cada uno de los empleados a su centro de labores se produce independientemente, de acuerdo a la distribución uniforme en el intervalo comprendido entre las 8.00 y 8.25 a.m.. Si 10 empleados llegaron al centro de labores después de las 8.10, ¿cuál es la probabilidad de que 4 de ellos hayan llegado entre las 8.15 y 8.20 a.m.?.

Rp. X: # de éxitos de 10, $X\sim B(10, p)$, $p=P[\text{éxito}]=P[8.15<T<8.20/T>8.1]=1/3$, $P[X=4]=0.2276$.

6. La demanda en miles metros de determinada tela que tiene una compañía textil, se distribuye según el modelo de uniforme en el intervalo [0,10]. Por cada metro de tela vendida se gana 4$, pero por cada metro de tela no vendida en la temporada se pierde $1.Calcular la producción que maximiza la utilidad esperada de la compañía.

Rp. X: demanda. X~U[0,10], sea K la producción Y: utilidad. Y=4K. si $x \geq K$, $Y=4x-1(K-x)$, si $x<K$. $E(Y)= (-5K^2/20)+4K$. $E(Y)$ es máximo si K=8

Normal

7. Suponga que la demanda mensual de un bien de consumo se distribuye normalmente con una media de 650 kilogramos y una desviación estándar de 100 kg.

a) ¿Qué probabilidad hay de que la demanda no supere los 500 kg.?.

b) ¿Qué cantidad del bien debe haber mensualmente a fin de satisfacer la demanda en el 89.8% de los meses?.

Rp. a) 0.0668. b) 777kg.

8. El porcentaje del ingreso ahorrado por las familias tiene distribución normal con una media del 10%.

a) Determine la desviación estándar, si el 2.28% de los ahorros son mayores que 12.4%.

b) ¿Qué porcentaje de familias ahorró más del 11.974% de sus ingresos?.

c) ¿Cuál es la probabilidad de que 3 familias de 5 tengas ahorros por más del 11.974% de sus ingresos?.

Rp. a) σ=1.2, b) 0..05 b) 0.001128.

9. Una pequeña ciudad es abastecida de agua cada dos días. El consumo en volumen de agua (cada dos días) tiene distribución normal.

a) Determine la media y varianza de la distribución si se sabe que el 0.62% del consumo es al menos de 22,500 litros y que el 1.79% del consumo es a lo más 17,900 litros.

b) Hallar la capacidad del tanque de agua de la pequeña ciudad para que sea sólo el 0.01 la probabilidad de que en el periodo de dos días el agua no sea suficiente para satisfacer toda la demanda.

Rp. a) μ = 20,000 litros, σ = 1,000 litros. b) 22,230 litros

10. Las calificaciones 400 alumnos en una prueba final de Estadística se distribuyen según el modelo de probabilidad normal con una media de 12.

a) Determine la desviación estándar si la nota mínima es 6 y la máxima es 18.

b) Si la nota aprobatoria es 11, ¿cuántos alumnos aprobaron el curso?.

c) ¿Qué nota como mínimo debería tener un alumno para estar ubicado en el quinto superior?

d) ¿Qué rango percentil tiene un alumno cuya nota es 14?, ¿indique su orden de mérito.

Rp. a) $\sigma \cong 2$, b) $276.6 \cong 277$, c) K tal que $P[X \geq K]=0.20$, $K=13.68$., d) $P[X \leq 14]=0.8413$, 15.87%

11. Los puntajes de una prueba de aptitud académica están distribuidos normalmente con una media de 60 y una desviación estándar de 10 puntos.

a) Si el 12.3% de los alumnos con mayor puntaje reciben el calificativo A y el 20% de los alumnos con menor nota reciben el calificativo C, calcular el mínimo puntaje que debe tener un alumno para recibir una A, y el máximo puntaje que debe tener un alumno para recibir una C.

b) Si el resto de los alumnos recibe el calificativo B y si el total de alumnos es igual a 90, ¿cuántos alumnos recibieron el calificativo de A, B y C?.

Rp. a) $\mu=60$, $\sigma=10$. b) A=71.6. C=51.6. c) A:11. B:61. C:18.

12. Las calificaciones de una prueba final de Matemática Básica tienen distribución normal con una media igual a 8. Si el 6.68% de los examinados tienen nota aprobatoria (mayor o igual a 11), ¿cómo debe modificarse cada nota para conseguir un 45% de aprobados?.

Rp. $\sigma=2$,. k=nota a añadir, $0.13=(11-(8+k))/2$, $k=2.74$.

13. Suponga que el ingreso familiar mensual (X) en una comunidad tiene distribución normal con media \$400 y desviación estándar \$50.

a) Si el décimo superior de los ingresos debe pagar un impuesto, ¿a partir de que ingreso familiar pagan el impuesto?.

b) Si el ahorro familiar está dada por la relación $Y=(1/4)X-50$, ¿cuál es la probabilidad de que el ahorro sea superior a \$75?

c) Si escogen dos familias al azar, ¿qué probabilidad hay de que una de ellas tenga ingresos mayores a \$498 y la otra menores que \$302?

Rp. a) K tal que, $P[X>K]=0.10$, $K=464$. b) $P[Y>75]=P[X>500]=0.0228$, c) $2\times 0.025\times 0.025$

14. Una pieza es considerada defectuosa y por lo tanto rechazada si su diámetro es mayor que 2.02 cm. O es menor que 1.98 cm.. Suponga que los diámetros tienen distribución normal con media de 2 cm. y desviación estándar de 0.01 cm.

a) Calcular la probabilidad de que una pieza sea rechazada

b) ¿Cuál es el número esperado de piezas rechazadas de un lote de 10,000 piezas?.

c) Si se escogen 4 piezas al azar, ¿cuál es la probabilidad de que 2 de ellos sean defectuosos?.

d) Se necesitan 4 piezas sin defecto para una máquina. Si estos se prueban uno a uno sin reposición, ¿cuál es la probabilidad de que la cuarta pieza buena sea la sexta probada?.

Rp. a) $p=0.0456$, b) $np=10{,}000\times 0.0456=456$, c) $C_2^4\, p^2(1-p)^2$, d) $C_3^5\ (1-p)^4\, p^2$.

15. Un gerente viaja diariamente en automóvil de su casa a su oficina y ha encontrado que el tiempo empleado en el viaje sigue una distribución normal con media de 35.5 minutos y desviación estándar de 3 minutos. Si sale de su casa todos los días a las 8.20 a.m. y debe estar en la oficina a las 9 a.m.

a) ¿Cuál es la probabilidad de que llegue tarde en un día determinado?

b) ¿Qué probabilidad hay de que llegue a tiempo a la oficina 3 días consecutivos?. Suponga independencia Rp. a) 0.0668, b) $(0.9332)^3$

16. Cierto líquido industrial contiene un porcentaje X por galón de un compuesto particular cuya distribución es normal con una media de 15% y una desviación estándar de 3%. El fabricante tiene una utilidad neta de \$0.15, si $9 < X < 21$, de \$0.10, si $21 \le X \le 27$, y una pérdida de \$.05, si $3 \le X \le 9$, calcular la utilidad esperada por galón.

Rp. valores: −0.05, 0.15, 0.10, prob. : 0.0228, 0.9544, 0.0228, espera: 0.14\$.

17. Un exportador recibe sacos de café de un quintal al mismo tiempo de dos proveedores A (Chanchamayo) y B (Quillabamaba). El 40% recibe de A y el resto de B. El porcentaje de granos con impurezas por saco es una variable aleatoria cuyo modelo de probabilidad es normal con media y desviación estándar respectivas de 6% y 2% para A, y de 8% y de 3% para B .

Si el exportador selecciona un saco al azar

a) ¿Qué probabilidad hay de que el porcentaje de granos con impurezas supere el 10%?

b) y encuentra que el porcentaje de granos con impurezas supere el 10%, ¿qué probabilidad hay de que provenga de Chanahmayo?.

Rp. a) $P= 0.4\times P[Z>2]+0.6\times P[Z>0.67]=0.4\times 0.0228+0.6\times 0.2514=0.15996$, b) $0.4\times 0.0228/ 0.15996$.

18. Una fábrica cuenta con 3 máquinas: A, B y C donde la máquina A produce diariamente el triple de B y ésta el doble de C. Además se sabe que el peso de los artículos producidos por A se distribuye exponencialmente con una media de 5 Kg., el peso de los artículos producidos por B se distribuyen uniformemente entre 3 Kg. y 8 Kg., mientras que el peso de los artículos producidos por C se distribuye normalmente con una media de 6 Kg. y una desviación estándar de 2Kg. Si se extrae al azar un artículo:

a) ¿Cuál es la probabilidad de que pese a lo más 5 Kg.?.

b) Si pesa más de 5 Kg., ¿Cuál es la probabilidad de que haya sido fabricado por la máquina A?.

Rp. E=peso ≤ 5, a) $P(E)=(6/9)\times 0.6321 +(2/9)\times 0.4 + (1/9)\times 0.3085=0.5446$,
b) $P(A/E^C)=P(A)P(E^C/A)/P(E^C)=(6/9)(0.3679)/0.4554$.

19. El monto de consumo que registra una cajera de un supermercado en una día cualquiera es una variable aleatoria que tiene distribución normal con media \$200 y desviación estándar \$50.

a) En este supermercado sólo el 5% de los clientes se considera un excelente cliente y por tanto como promoción puede recibir un 10% de descuento, ¿a partir de que consumo un cliente se beneficiará de la promoción?.
b) Actualmente el 30% de los clientes tiene un consumo considerado como mínimo. La empresa considera que en base a la promoción en unos meses sólo el 20% de los clientes consumirá debajo de ese monto, ¿cuánto dinero adicional tendrá que gastar cada cliente para que esto se cumpla?.

Rp. $X \sim N(200, (50)^2)$, a) k tal que $P[X \geq k]=0.05$, $k=282.25$, b) $P[X<c]=0.30$, $c=174$. Sea d dinero adicional, $Y=X+d$, $Y \sim N(200+d, (50)^2)$, hallar d tal que, $P[Y<174]=0.20$, $d=16$.

20. Suponga que el costo de consumo por persona en un restaurante se distribuye normalmente con una desviación estándar igual a \$5. Si se sabe que el 15.87% de los clientes han pagado más de \$15 y que 112 personas pagaron menos de \$7.1, ¿cuántas personas comieron en el restaurante?.

Rp. μ =\$10, 400 personas.

Propiedad reproductiva de la normal

21. Los pesos de los posibles usuarios de un ascensor constituyen una población cuya distribución es normal con una media de 70 kg. y una desviación estándar de 10 kg.

a) ¿Qué peso máximo debería poder soportar el ascensor de modo que sólo el 1% de las ocasiones el peso de 4 personas supere ese peso máximo?.
b) Si el ascensor admite como peso máximo 585 kg., ¿cuántas personas a la vez pueden entrar al ascensor de manera que sea 0.0668 la probabilidad de que el peso no supere el máximo permitido?.

Rp. a) $k = 326.6$ kg. , b) $n=9$

22. Se ha determinado que los salarios, en dólares, de las parejas de esposos son independientes y que la distribución es $N(350, 50^2)$ para los hombres y $N(250,35^2)$ para las mujeres

a) ¿Cuál es la probabilidad de que el ingreso familiar sea superior a \$720.?
b) Si se escoge al azar una pareja de esposos, ¿cuál es la probabilidad de que el sueldo del esposo sea mayor que el sueldo de la esposa?
c) Si se escoge al azar una pareja de esposos, ¿cuál es la probabilidad de que cada ingreso sea más de \$300?.

Rp. a) 0.0244, b) $P[H>M]=P[H-M>0]= P[Z>(0-100)/61.033]= 0.9495$
c) $P[H>300 \wedge M>300]=P[H>300]\times P[M>300]=P[Z>-1]\times P[Z>1.43]= 0.0643$.

23. Suponga que el peso de las botellas vacías de gaseosas tienen un peso con distribución normal de media 0.4 kg. y desviación estándar 0.01 kg. El peso del líquido que se depositan en las botellas tiene una distribución normal con media 0.7 kg y desviación estándar 0.05 kg. Los pesos de las cajas vacías donde se

colocan las botellas tienen una distribución normal de media 2 kg y desviación estándar 0.05 kg.. Si cada caja contiene 12 botellas llenas de tal gaseosa;

a) ¿Cuál es la probabilidad de que el peso de una caja de 12 botellas llenas pese menos de 15 kg?.

b) Si se tienen 10 cajas llenas, ¿cuál es la probabilidad de que 8 de ellas pesen menos de 15 kg?.

Rp. a) X=peso total de una caja~$N(15.2, (0.1836)^2)$, $p=P[X<15]=0.1379$. b) $C_8^{10}p^8(1-p)^2$.

24. La temperatura de un objeto es una variable aleatoria con modelo de probabilidad $f(x)=(1/\sqrt{\pi})e^{-x^2}$, $-\infty<x<+\infty$. Cierto tipo de artículo tiene un precio de venta fijo de 25 unidades monetarias (u.m.) y un costo variable que es $N(5,1)$. ¿Cuál es la probabilidad de que la venta de 6 artículos origine una utilidad neta total mayor que 125 u.m.?.

Rp. 0.0207.

Gamma y exponencial

25. La distribución de la duración en meses de cierto tipo de objeto es exponencial con parámetro β, ¿cuál es el valor de β si se sabe que hay una probabilidad de 0.7 de que uno de estos objetos tenga una duración a lo más de 6 meses?.

Rp. $\beta = 0.2$.

26. Suponga que el tiempo de vida útil de un modelo de computadora es una variable aleatoria con distribución exponencial cuya media es 10 meses.

a) Si el costo del montaje de cada computadora es \$660 y la venta es \$1000, determinar la utilidad esperada por cada computadora sabiendo que el distribuidor cambia por otro nuevo si dura menos de 5 meses.

b) Una empresa adquiere 10 de tales computadoras, ¿cuál es la probabilidad de que al menos 1 dure al menos 12 meses?

Rp. a) $-320(1-e^{-0.5})+340(e^{-0.5})=80.3$. b) $X\sim B(10,p)$, $p=0.3$. $P[X\geq 1]=0.972$.

27. La vida útil de una batería en años es una variable aleatoria X con la función de densidad:

$$f(x)=0.2e^{-0.2x}, \qquad x\geq 0.$$

El fabricante ofrece una garantía de un año. Si la batería falla en ese período se reemplaza por otra, a lo más una sola vez. Si el costo de fabricación es de \$20 por cada batería y se vende a \$50 cada una,

a) ¿Cuál es la utilidad esperada del fabricante?

b) ¿Cuál debe ser el tiempo de garantía que el fabricante debe ofrecer para que sólo devuelva el 5% de las baterías producidas?.

Rp. a) $30(e^{-0.2})+10(1-e^{-0.2})=26.37$. b) $0.05=P[X\leq k]$. $k=0.2564$

28. El 90% de los objetos de un proceso de producción son clasificados como buenos y el resto como defectuosos. Si la vida útil (en años) de los objetos

defectuosos y buenos se distribuyen ambos exponencialmente con medias respectivas 0.2 años y 0.5 años,

a) ¿Cuál es la probabilidad de que un objeto seleccionado al azar dure a lo más un año?.

b) Si el objeto seleccionado al azar dura a lo más un año, ¿cuál es la probabilidad de que haya sido seleccionado de la producción defectuosa?.

Rp. X: duracion. D: defectuoso, $E=\{X\leq 1\}$. a) $P(E)=P(D)P(E/D)+P(D^C)P(D^C/E)$. $P(E)=0.1\times(1-e^{-5})+0.9\times(1-e^{-2})=0.8775$, b) $P(D/E)=0.0993/0.8775$.

29. La **confiabilidad** de una componente (o un sistema) durante un periodo t es la probabilidad de que su tiempo de falla exceda t. Esto es, la confiabilidad es $R(t)=P[T>t]$.

Un fabricante ha determinado que el tiempo de falla en meses de cierto modelo de motor tiene la función de densidad

$$f(t)=0.025e^{-0.025t}, \quad t\geq 0$$

a) Determinar la confiabilidad del motor para un periodo de 4 años.

b) Si el tiempo de falla del motor excede el tiempo promedio de falla, ¿cuál es la probabilidad de que exceda un año más?

c) Hallar el periodo en el cual el motor tiene una confiabilidad igual 0.325.

Rp. a) $R(48)=P[T>48]=e^{-1.2}=0.3$, b) $P[X>40+12/X>40]=P[X>12]=e^{-0.3}$, c) $e^{-k(0.025)}=0.325$, $k=44.96\cong 45$.

30. La distribución de la duración de los fusibles que produce un fabricante tanto por el proceso I como por el proceso II es exponencial con medias respectivas de 100 y 150 horas. Los costos de fabricación por fusible y para cada proceso son \$0.5 y \$0.8 respectivamente. Sin embargo, si cualquier fusible dura menos de 200 horas, se carga una pérdida de \$0.2 en contra del fabricante, ¿en cuánto deberían venderse los fusibles de cada tipo si se espera una utilidad de \$1 en cada caso?. Rp. \$1.67 y \$1.95.

31. El número de clientes que usan tarjetas de crédito pasan por la caja 25 de un supermercado a razón de 15 por hora, siguiendo un proceso de Poisson, ¿qué probabilidad hay de que

a) transcurra más de 10 minutos para que pase el primer cliente por la caja 25?,

b) el 2do cliente pase por la caja 25 cinco minutos después de puesta en marcha la caja?.

Rp. a) $T\sim\Gamma(\alpha=1,\beta=5/2)$, $P[T>10]=P[X=0]=e^{-5/2}$, b) $T\sim\Gamma(\alpha=2,\beta=1/4)$, $P[T>5]=P[X\leq 1]=e^{-(5/4)}[1+5/4]$.

32. Se ha determinado que en promedio la demanda por un artículo en particular en una bodega era 1 cada 20 minutos. Considerando que el número de artículos demandados se distribuye según la ley de Poisson, halle la probabilidad de que el quinto artículo sea demandado después de dos horas que se abrió la bodega.

Rp. $P[T>120]=P[X\leq 4]=e^{-6}6^0/0!+\ldots+e^{-6}6^4/4!=0.285$. $T\sim\Gamma(\alpha=5,\beta=1/20)$, en minutos $X\sim P(6)$

7.3 Teorema del límite central

Uno de los teoremas sobre límite que justifica la importancia que tiene la distribución normal es denominado *teorema del límite central*. Este teorema nos permite aproximar a la distribución normal sumas finitas de variables aleatorias independientes que pueden tener cualquier clase de distribución con medias y varianzas conocidas.

Se dice que n variables aleatorias $X_1, X_2, \cdots, X_n$ tienen la misma distribución de probabilidad con media μ y varianza σ^2 si, tienen la misma función de probabilidad en el caso discreto, o la misma función de densidad en el caso continuo, y si además,

$$E(X_i) = \mu, \quad Var(X_i) = \sigma^2 \text{ para } i = 1,2,3,\ldots,n.$$

Si estas n variables aleatorias son independientes, entonces, la variable aleatoria $Y_n = X_1 + X_2 + \cdots + X_n$, tiene:

media $E(Y_n)$ o μ_{Y_n}

y Varianza $Var(Y_n)$ o $\sigma^2_{Y_n} = n\sigma^2$

TEOREMA 7.12

Si $X_1, X_2, \cdots, X_n$ son n variables aleatorias independientes igualmente distribuidas con media μ y varianza σ^2, entonces, la distribución de la variable aleatoria:

$$Z_n = \frac{Y_n - \mu_{Y_n}}{\sigma_{Y_n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

tiende a la distribución normal $N(0,1)$, cuando n tiende al infinito.

Esto es,
$$\lim_{n\to\infty} P[Z_n \le b] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b} e^{-t^2/2} dt = \Phi(b).$$

NOTAS.

1. La distribución de la variable aleatoria $Y_n = X_1 + X_2 + \cdots + X_n$ es aproximadamente normal con media μ_{Y_n}, y con desviación estándar $\sigma_{Y_n} = \sigma\sqrt{n}$. Luego,

$$P[Y_n \leq b] \cong P\left[Z_n \leq \frac{b - n\mu}{\sigma\sqrt{n}}\right]$$

La ***aproximación es buena si*** $n \geq 30$.

2. Si la variable aleatoria $Y_n = X_1 + X_2 + \cdots + X_n$ es discreta y con valores enteros consecutivos, entonces,

$$P[Y_n = k] \cong P[k - 0.5 \leq Y_n \leq k + 0.5]$$

y
$$P[a \leq Y_n \leq b] \cong P[a - 0.5 \leq Y_n \leq b + 0.5].$$

El ajuste de 0.5 se denomina ***factor de corrección por continuidad***.

EJEMPLO 7.31

Sean $X_1, X_2, \cdots, X_{50}$, 50 variables aleatorias independientes cada una distribuida según la función de probabilidad de la figura 7.13. Calcular la probabilidad

$$P[X_1 + X_2 + ... + X_{50} > 55].$$

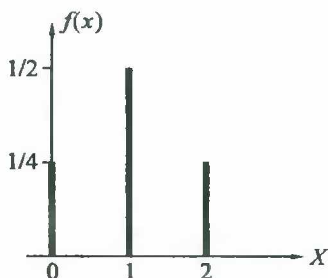


Fig. 7.13

SOLUCION.

De la figura 7.13 se deduce que cada variable aleatoria X_i tiene la distribución de probabilidad dada por la tabla que sigue:

su media y su varianza son respectivamente:

$$E(X_i)=0\left(\frac{1}{4}\right)+1\left(\frac{1}{2}\right)+2\left(\frac{1}{4}\right)=1.$$

$$Var(X_i)=E\left(X_i^2\right)-[E(X_i)]^2=0^2\left(\frac{1}{4}\right)+1^2\left(\frac{1}{2}\right)+2^2\left(\frac{1}{4}\right)-1^2=\frac{1}{2}$$

x	0	1	2
$f(x)=P[X=x]$	1/4	1/2	1/4

Si la variable aleatoria, $Y_{50}=X_1+X_2+\cdots+X_{50}$, entonces,

$$E(Y_{50})=E(X_1+X_2+...+X_{50})=\sum_{i=1}^{50}E(X_i)=50(1)=50$$

$$Var(Y_{50})=Var\left(\sum_{i=1}^{50}X_i\right)=\sum_{i=1}^{50}Var(X_i)=50(\frac{1}{2})=25$$

Por el teorema del límite central, la variable aleatoria:

$$Z=\frac{Y_{50}-E(Y_{50})}{\sqrt{Var(Y_{50})}}=\frac{Y_{50}-50}{5}$$

tiene distribución aproximadamente normal $N(0,1)$. Entonces,

$$P[Y_{50}>55]=1-P[Y_{50}\le 55]\cong 1-P[Y_{50}\le 55.5]=1-P\left[Z\le\frac{55.5-50}{5}\right].$$

Luego, $P[Y_{50}>55]\cong 1-P[Z\le 1.1]=1-0.8643=0.1357.$

EJEMPLO 7.32

El costo de un producto consiste de un valor fijo de \$5 y de un valor variable que es el 5% de la garantía. Suponga que la garantía tiene distribución uniforme entre \$0 y \$200. Si el producto se vende en \$15

a) Calcule la probabilidad de que la utilidad de 50 unidades del producto supere los \$290.

b) ¿Cuántas unidades del producto serían necesarias vender para que con probabilidad 0.5 la utilidad sea al menos \$300?

SOLUCION.

Sea Y_i la utilidad y X_i el costo variable de la i-ésima unidad del producto.

La variable $X_i \sim U[0,200]$, con media y varianza respectivas:

$$E(X_i)=\frac{0+200}{2}=100,\ \ V(X_i)=\frac{(200-0)^2}{12}=\frac{200^2}{12}.$$

La utilidad: $\qquad Y_i=15-(5+0.05X_i)=10-0.05X_i$.

Entonces, la media y la varianza de Y_i son respectivamente:

$$E(Y_i)=5 \ \text{ y } \ V(Y_i)=\frac{100}{12}.$$

a) Sea $U_{50}=\sum_{i=1}^{50} Y_i$ la utilidad que producen 50 unidades del producto. Entonces, la variable aleatoria estándar:

$$Z_{50}=\frac{U_{50}-50(5)}{\sqrt{50(100/12)}}=\frac{U_{50}-250}{20.41}\cong N(0,1).$$

Luego, $\qquad P[U_{50}>290]\cong P\left[Z_{50}>\frac{290-250}{20.41}\right]=P[Z_{50}>1.96]=0.025.$

b) Sea $U_n=\sum_{i=1}^{n} Y_i$ la utilidad que producen n unidades del producto. entonces, la variable aleatoria estándar:

$$Z_n=\frac{U_n-5n}{\sqrt{100n/12}}\cong N(0,1).$$

Se debe determinar el valor de n tal que $P[U_n\geq 300]=0.5$. Entonces,

$$P[U_n\geq 300]\cong P\left[Z_n\geq\frac{300-5n}{\sqrt{n(100/12)}}\right]=0.5$$

$$\frac{300-5n}{\sqrt{n(100/12)}}=0$$

De donde resulta $\qquad n=60.$

EJEMPLO 7.33

Una máquina automática envasa un bien de consumo diario en bolsas cuyo peso total tiene una media de 25 gramos y una desviación estándar de 1 gramo. Una segunda máquina empaca estas bolsas en paquetes de 12 docenas. Con el fin de verificar si el número de unidades es realmente 144 en cada paquete, se escoge uno al azar y se adopta la siguiente regla de decisión: si el peso del paquete está entre 3576 y 3624 gramos se acepta que el paquete tiene 144 bolsas, de otro modo se rechaza. Calcular la probabilidad de que,

a) Un paquete que realmente tiene 144 bolsas sea considerado como si no los tuviera.
b) Un paquete que tiene 142 bolsas sea considerado como si tuviera 144 bolsas.

SOLUCION.

Sea X_i el peso de la bolsa *i*, $i = 1, 2, ..., 144$. La distribución de cada X_i es de forma desconocida, pero tienen,

$$E(X_i) = 25\text{, y}\quad V(X_i) = 1.$$

a) Sea la variable aleatoria: $Y_{144} = \sum_{i=1}^{144} X_i$ el peso de un paquete de 144 bolsas,

entonces: $$E(Y_{144}) = n\mu = 144 \times 25gr. = 3600gr.$$

$$Var(Y_{144}) = n\sigma^2 = 144 \times 1gr^2 = 144gr^2.$$

Por el teorema central del límite, la variable aleatoria:

$$Z = \frac{Y_{144} - E(Y_{144})}{\sqrt{Var(Y_{144})}} = \frac{Y_{144} - 3600}{12}$$

tiene distribución aproximadamente normal $N(0,1)$. Luego,

$$P[Y_{144} < 3576 \text{ o } Y_{144} > 3624] = 1 - P[3576 \le Y_{144} \le 3624] = 0.0456$$

donde

$$P[3576 \le Y_{144} \le 3624] = P\left[\frac{3576 - 3600}{12} \le Z \le \frac{3624 - 3600}{12}\right]$$

$$= P[-2 \le Z \le 2] = 0.9544.$$

b) Sea la variable aleatoria: $Y_{142} = \sum_{i=1}^{142} X_i$ el peso de un paquete de 142 bolsas,

entonces,

$$E(Y_{142}) = n\mu = 142 \times 25gr. = 3550gr.$$

$$Var(Y_{142}) = n\sigma^2 = 142 \times 1gr^2 = 142gr^2.$$

Por el teorema central del límite, la variable aleatoria:

$$Z = \frac{Y_{144} - E(Y_{144})}{\sqrt{Var(Y_{144})}} = \frac{Y_{144} - 3550}{11.916}$$

tiene distribución aproximadamente normal $N(0,1)$. Luego,

$$P[3576 \leq Y_{142} \leq 3624] = P\left[\frac{3576-3550}{11.916} \leq Z \leq \frac{3624-3550}{11.916}\right]$$

$$= P[2.18 \leq Z \leq 6.21] = 0.0146.$$

EJEMPLO 7.34

Suponga que la vida útil de una componente eléctrica de trabajo continuado tiene distribución exponencial con un promedio de 100 horas. Tan pronto como se deteriora la componente, es reemplazada por otra para continuar con el trabajo.

a) Calcular la probabilidad de que durante 209.5 días, se necesiten más de 36 de estas componentes.

b) ¿Cuántas de estas componentes se necesitan para que duren al menos 4.536 horas con una probabilidad de 0.9901?.

SOLUCION.

Sea la variable aleatoria X_i la duración de la componente i, $i = 1, 2, 3, ..., n$. Las variables aleatorias X_i son independientes y están distribuidas cada una exponencialmente con media $\mu = 100$ horas y con desviación estándar igual a la media esto es, $\sigma = 100$ horas. El tiempo total que duran las n componentes es la variable aleatoria:

$$Y_n = \sum_{i=1}^{n} X_i$$

entonces, $\quad E(Y_n) = n\mu = 100n$ y $Var(Y_n) = n\sigma^2 = 100^2 n$

Por el teorema central del límite, la variable aleatoria :

$$Z = \frac{Y_n - E(Y_n)}{\sqrt{Var(Y_n)}} = \frac{Y_n - 100n}{100\sqrt{n}}$$

tiene distribución aproximadamente normal $N(0,1)$.

a) Se necesitan más de 36 componentes durante 209.5 días, si la vida útil total es menor que 5,028 horas (209.5 por 24 horas). Luego, hay que calcular:

$$P[Y_{36} < 5{,}028] = P\left[Z_{36} < \frac{5{,}028 - 36(100)}{100\sqrt{36}}\right] = P[Z_{36} < 2.38] = 0.9913$$

b) Se debe calcular n tal que $P[Y_n \geq 4{,}536] = 0.9901$, luego,

$$0.0099 = P[Y_n < 4536] \cong P\left[Z_n < \frac{4{,}536 - 100n}{100\sqrt{n}}\right]$$

$$-2.33 = \frac{4{,}536 - 100n}{100\sqrt{n}}$$

$$100n - 233\sqrt{n} - 4{,}536 = 0$$

de donde resulta, $n = 64$.

7.3.1 Aproximaciones a la distribución normal

7.3.1.1 Aproximación de la binomial a la normal

Sean $X_1, X_2, ..., X_n$, n variables aleatorias independientes e idénticamente distribuidas según el modelo Bernoullí con parámetro p. Cada variable X_i, tiene media $\mu_{X_i} = p$ y varianza $\sigma^2_{X_i} = pq$, siendo $q = 1 - p$. Luego, la variable aleatoria:

$$X = \sum_{i=1}^{n} X_i$$

tiene distribución binomial con media $\mu_X = np$ y con varianza $\sigma^2_X = npq$.

Por el teorema central del límite, la variable aleatoria estándar:

$$Z=\frac{\sum_{i=1}^{n} X_i - n\mu_{X_i}}{\sqrt{n\sigma^2_{X_i}}}=\frac{X-np}{\sqrt{npq}}$$

tiene aproximadamente distribución normal $N(0,1)$, cuando n es suficientemente grande.

Luego, si X es una variable aleatoria binomial con media np y varianza npq, entonces, cuando $n \to +\infty$ la distribución de la variable aleatoria:

$$Z=\frac{X-np}{\sqrt{npq}}$$

es aproximadamente la distribución normal $N(0,1)$.

La aproximación es buena si $n \geq 30$. Si $n < 30$, la aproximación es buena si p es cercano a 0.5. Cuánto más se aleja p de 0.5, se requiere n cada vez más grande para tener una aproximación aceptable.

Además, se puede tener mayor precisión en la aproximación si se utiliza **la corrección por continuidad** de los valores discretos de la variable, sumando o restando 0.5 como sigue:

$$P[a \leq X \leq b] \cong P[a-0.5 \leq X \leq b+0.5] \cong \Phi\left[\frac{b+0.5-np}{\sqrt{npq}}\right]-\Phi\left[\frac{a-0.5-np}{\sqrt{npq}}\right]$$

y $$P[X=k]=P[k-0.5 \leq X \leq k+0.5]$$

EJEMPLO 7.35

El gerente de ventas de TV cable estima en 30% la proporción de clientes morosos. Si se selecciona al azar una muestra de 200 clientes

a) Halle la probabilidad de que más de 50 de ellos sean morosos
b) ¿Cuál es la probabilidad de que exactamente 70 clientes sean morosos?.
c) Calcule la probabilidad de que el número de clientes morosos de la muestra difiera del promedio en no más de 14 clientes, si en verdad es 0.2 la proporción de clientes morosos.

SOLUCION

Sea X el número de clientes morosos en la muestra. Entonces X tiene distribución binomial $B(n,p)$, donde $n=200$, $p=0.3$.

a) La probabilidad de que **más de 50 clientes de la muestra** sean morosos es:

$$P[X > 50] = P[X \geq 51] = \sum_{k=51}^{200} C_k^{200} (0.3)^k (0.7)^{200-k}$$

Utilizando la **aproximación** a la normal $N(\mu, \sigma^2)$, de la variable binomial X se tiene:

$$\mu = np = 200 \times 0.3 = 60 \qquad \sigma = \sqrt{npq} = \sqrt{200 \times 0.3 \times 0.7} = 6.48.,$$

y que $Z = \dfrac{X - 60}{6.48}$ es aproximadamente distribuida como normal $N(0,1)$. Luego,

$$P[X > 50] \cong P[X \geq 50.5] = P\left[Z \geq \frac{50.5 - 60}{6.48}\right] = P[Z \geq -1.47] = 0.9292$$

b) $P[X = 70] = C_{70}^{200} (0.3)^{70} (0.7)^{130}$

Utilizando la aproximación a la normal de la variable binomial X se tiene que $Z = \dfrac{X - 60}{6.48}$ es aproximadamente distribuida como normal $N(0,1)$. Por tanto,

$$P[X = 70] \cong P[69.5 \leq X \leq 70.5] = P[1.47 \leq Z \leq 1.62] = 0.0182.$$

c) En este caso, $Z = \dfrac{X - np}{\sqrt{npq}} = \dfrac{X - 200 \times 0.2}{\sqrt{200 \times 0.2 \times 0.8}} = \dfrac{X - 40}{5.657}$ es aproximadamente distribuida como normal $N(0.1)$. Por tanto.

$$P[|X - \mu| \leq 14] = P\left[\left|\frac{X - 40}{5.657}\right| \leq \frac{14}{5.657}\right] = P[|Z| \leq 2.47] = 0.9864$$

7.3.1.2 Aproximación de la hipergeométrica a la normal

Sea X la variable aleatoria hipergeométrica. esto es, el número de éxitos que se encuentran en una muestra de tamaño n escogida al azar una a una sin reposición de N elementos de los cuales r son clasificados como éxitos y el resto $N - r$ como fracasos. La media y la varianza de la variable son respectivamente:

$$\mu = np \quad \text{y} \quad \sigma^2 = npq\frac{N - n}{N - 1} \qquad \text{donde } p = r/N, \text{ y } q = 1 - p.$$

Cuando $n \to +\infty$

$$Z = \frac{X - np}{\sqrt{npq\,\dfrac{N-n}{N-1}}}$$

es aproximadamente la distribución $N(0,1)$.

EJEMPLO 7.36

Los 1000 sacos de café de un quintal cada uno, producto de la última cosecha en Chanchamayo, se envían a una empresa exportadora para su venta a USA. La empresa exportará el producto si una muestra aleatoria de 80 sacos (revisados uno por uno) revela a lo más 12.5% de sacos que no cumplen las especificaciones, en caso contrario no lo exportará. Calcular la probabilidad de no exportar el producto cuando realmente el porcentaje de sacos que no cumplen las especificaciones es 10% del total.

SOLUCION.

Sea X la variable aleatoria que denota el número de sacos de café de un quintal que no cumplen las especificaciones en la muestra de $n = 80$ escogida de la **población finita** de 1000 sacos de los cuales 100 no cumplen las especificaciones.. Entonces, X tiene distribución hipergeométrica $H(1000,100,80)$.

El producto será exportado si el número de sacos que no cumplen las especificaciones en la muestra es 10 o menos (a lo más el 12.5% de 80). Luego la probabilidad de que el producto no sea exportado es:

$$P[X > 10] = 1 - P[X \le 10] = 1 - \sum_{k=0}^{10} \frac{C_k^{100} C_{80-k}^{900}}{C_{80}^{1000}}$$

Para calcular la suma, utilizamos la **aproximación normal** $X \cong N(\mu, \sigma^2)$, de la variable hipergeométrica X donde, $\mu = np = 80 \times \dfrac{100}{1000} = 8$

$$\sigma = \sqrt{np(1-p)\frac{N-n}{N-1}} = \sqrt{80 \times 0.1 \times 0.9 \times \frac{1000-100}{1000-1}} = 2.547$$

Por el teorema central del límite la variable $Z = \dfrac{X - 8}{2.547}$ tiene distribución aproximadamente normal $N(0,1)$.

Por tanto,

$$\sum_{k=0}^{10}\frac{C_k^{100}C_{80-k}^{900}}{C_{80}^{1000}} \cong P\left[Z \le \frac{10.5-8}{2.547}\right] = P[Z \le 0.98] = 0.8365$$

$$P[X > 10] \cong 1 - 0.8365 = 0.1634.$$

Observar que si no se hace corrección por continuidad se obtiene:

$$P[X > 10] = 1 - P[X \le 10] \cong 1 - P[Z \le 0.79] = 1 - 0.7852 = 0.2148.$$

7.3.1.3 Aproximación de la Poisson a la normal

Sean $X_1, X_2, ..., X_n$, n, variables aleatorias independientes e idénticamente distribuidas como Poisson cada una con promedio λ en un intervalo dado. Es decir en el mismo intervalo cada X_i, ocurre con media y varianza iguales a λ. Entonces, para n suficientemente grande la variable:

$$X = \sum_{i=1}^{n} X_i$$

tiene distribución de Poisson con media y varianzas iguales a $n\lambda$.

En consecuencia, por el teorema del límite central, la variable aleatoria:

$$Z = \frac{X - n\lambda}{\sqrt{n\lambda}}$$

tiene aproximadamente distribución $N(0,1)$, cuando n es suficientemente grande.

La aproximación es buena si $n\lambda > 5$.

EJEMPLO 7.37

El promedio del número de accidentes de trabajo en una fábrica es 2 cada semana. Calcule la probabilidad de que se produzcan al menos 115 accidentes durante 50 semanas.

SOLUCION

Sea X_i; # de accidentes por semana con un promedio de $\lambda = 2$. Entonces,

$$X = \sum_{i=1}^{50} X_i \text{ , el número de accidentes en 50 semanas.}$$

tiene distribución de Poisson con media y varianzas iguales a

$$n\lambda = 50 \times 2 = 100 .$$

Por el teorema del límite central, la variable aleatoria:

$$Z = \frac{X - n\lambda}{\sqrt{n\lambda}} = \frac{X - 100}{10}$$

tiene aproximadamente distribución $N(0,1)$

$$P[X \geq 115] \cong P[X \geq 114.5] = P[Z \geq 1.45] = 0.0735.$$

7.3.1.4 Aproximación de la Chi-cuadrado a la normal

Si la variable aleatoria $X \sim \chi^2(r)$, entonces, para r suficientemente grande, la variable aleatoria $\sqrt{2X}$ tiene aproximadamente distribución $N\left(\sqrt{2r-1},1\right)$. En consecuencia, $Z = \sqrt{2X} - \sqrt{2r-1} \sim N(0,1)$, y

$$P[X \leq a] = = P\left[Z \leq \sqrt{2a} - \sqrt{2r-1}\right] \cong \Phi\left(\sqrt{2a} - \sqrt{2r-1}\right)$$

el valor de Φ se obtiene de la tabla de la distribución normal.

EJEMPLO 7.38

a) Si $X \sim \chi^2(200)$, calcular $P[160 \leq X \leq 240]$.

b) Si $X \sim \chi^2(100)$, calcular a tal que $P[X \leq a] = 0.995$.

SOLUCION.

a) Si $X \sim \chi^2(200)$, entonces, $Z = \sqrt{2X} - \sqrt{2 \times 200 - 1} \sim N(0,1)$. Luego

$$P[160 \leq X \leq 240] \cong P[\sqrt{2\times 160} - \sqrt{399} \leq Z \leq \sqrt{2 \times 240} - \sqrt{399}]$$

$$\cong P[-2.09 \leq Z \leq 1.93] = 0.9732 - (1 - 0.9817) = 0.9549.$$

b) Si $X \sim \chi^2(100)$, entonces, $Z = \sqrt{2X} - \sqrt{2 \times 100 - 1} \sim N(0,1)$. Luego,

$$0.995 = P[X \leq a] = P[Z \leq \sqrt{2a} - \sqrt{2(100)-1}] \cong P[Z \leq \sqrt{2a} - 14.11]$$

entonces, $\sqrt{2a} - 14.11 \cong 2.575$, de donde resulta $a = 139.1$

EJERCICIOS

1. Se ha determinado que la vida útil de cierta marca de llantas radiales tiene una media de 38,000 Km. y una desviación estándar de 3,000 Km. Si una empresa de transporte adquiere 36 de estas llantas, ¿cuál es la probabilidad de que rindan al menos 1,318,500 km. en conjunto?.

Rp. 0.9970

2. El número de ingresos a internet que ocurre diariamente en determinada computadora personal, es una variable aleatoria X con distribución de probabilidades:

X	0	2	4
$P[X=x]$	0.125	0.75	0.125

a) Determine la distribución de probabilidades del número de ingresos de 30 días.

b) Calcule la probabilidad de que ocurran entre 50 y 70 ingresos a internet en tal computadora en un periodo de 30 días

Rp. $\mu=2$, $\sigma^2=1$, $Y_{30}=\sum X$, demanda 30 días, a) $Y_{30}\cong N(60,30)$, b) $P[50\leq Y_{30}\leq 70]=0.9328$ (sin corregir).

3. Suponga que en cierta región el ingreso mensual por familia en miles de dólares es una variable aleatoria X con función de densidad:

$$f(x)=\begin{cases} 0.25x & \text{si } 0\leq x\leq 2 \\ 1-0.25x & \text{si } 2\leq x\leq 4 \end{cases}$$

Considerando una muestra aleatoria de 100 ingresos

a) ¿Cuál es la probabilidad de que el total de ingresos de la muestra supere los 167 mil dólares?.

b) Calcular aproximadamente la probabilidad de que más de 20 ingresos de la muestra sean mayores que 3 mil dólares?

Rp. $E(X)=2$, $V(X)=4.6667$. a) $Y_{100}=\sum X_i$ total de 100 ingresos, $P[Y_{100}>167]=0.063$.
b) W:# de ingresos mayores 3 mil dólares, $W\sim B(100,p)$, $p=P[X>3]=0.125$,
$P[W>20]=P[Z\geq(20.5-12.5)/3.307]=0.0078$.

4. La demanda diaria de un producto es una variable aleatoria X con distribución de probabilidad:

x	1	2	3	4	5
$P[X=x]$	1/10	2/10	3/10	3/10	1/10

a) ¿Cuál es la probabilidad de que la demanda de 30 días supere las 103 unidades?

b) Cada día se ponen a la venta 4 unidades del producto. Cada unidad se compra a \$5 y se vende \$10 unidad. Cuando una unidad del producto no se vende durante el día se descarta con una pérdida adicional de \$0.5, calcular la probabilidad de que la utilidad que se genere en 36 días supere los \$495.

Rp. μ=3.1, σ^2=1.29, $Y_{30}=\sum X$, demanda de 30 días, $Y_{30}\cong N(93,38.7)$, $P[Y_{30}>103]$=0.0537 (sin corregir). U=Utilidades: −11.5, −1, 9.5, 20, 20, $E(U)$=9.5, $V(U)$=110.25, W suma de 36 utilidades, $W\cong N(342, 63^2)$, $P[W>495]\cong$ 0.0075.

5. Se desea fabricar cables de fibras de nylon de manera que puedan resistir sin romperse al menos 466 Kg.. Si cada fibra tiene una resistencia media de 10 Kg., y una varianza de 0.75 kg^2., ¿Con cuántas fibras se debe formar el cable de manera que cumpla las exigencias con probabilidad 0.99?

Rp. n=? / $P[Y_n\geq 466]$=0.99, $P[Z\geq(466-10n)/(0.75n)^{1/2}]$=0.99, n=48.

6. Suponga que cada bolsa de cemento mezclado con arena y piedra chancada llena en promedio 4 mts^2 de techo con una desviación estándar de 1.1 mts^2. Asumiendo independencia en el rendimiento de cada bolsa de cemento, si se va a llenar un techo de 188 mts^2, ¿cuántas bolsas de cemento como mínimo serán necesarias para que con probabilidad 0.85 se logre llenar el techo?

Rp. n=? / $P[Y_n\geq 188]$=0.99, $P[Z\geq(188-4n)/1.1(n)^{1/2}]$=0.85, n=49.

7. El tiempo que demora un operario en ensamblar un objeto es una variable aleatoria X, cuya distribución tiene una media de 30 minutos y una desviación estándar de 2 minutos. El objeto totalmente terminado requiere un tiempo de $5+X$ minutos. Si el operario tiene que entregar 36 objetos totalmente terminados, calcular la probabilidad de que utilice al menos 20.5 horas.

Rp. 0.9938.

8. El tiempo que demora una persona en hablar por teléfono es una variable aleatoria con media de 3 minutos y una desviación estándar de 0.5 minutos. Si el costo por llamada tiene un valor fijo de \$0.8 más un costo variable de \$0.05 por minuto, calcular la probabilidad de que el costo al realizar 36 llamadas sea mayor que \$85.

Rp. X: tiempo, C: costo $C=0.8+0.5X$, E(C)=2.3, V(C)=0.0625.: Y_{36}:costo 36 llamadas, $Y_{36}\cong N(82.2, 2.25)$, $P[Y_{36}>\$85]$= 0.0307

9. Un supermercado produce un pan especial cuyo peso X debe tener una media de 100 gramos y una desviación estándar de 5 gramos. Si el pan tiene más de 100 gramos, el costo por la diferencia por cada pan en soles está dado por

$$C=0.0125X-1.00$$

Si se producen 200 panes por turno, ¿cuál es la probabilidad de que el costo total por la diferencia supere los S/48?

Rp. $E(C_i)=0.25$, $V(C_i)=0.00391$, $Y_{200}=\sum C_i$ el total de 200 costos.
$P[Y_{200}>48]=P[Z>(48-50)/0.8839]=0.9881$

10. ¿Cuál es la probabilidad de que en una muestra de 80 alumnos existan a lo más dos con enfermedad E si se sabe que en toda la población,

el 70% son hombres
el 30% son mujeres
el 5% de los hombres padecen la enfermedad E
el 2% de las mujeres padecen la enfermedad E?

Rp. $P(E)=0.05\times0.7+0.02\times0.3=0.041$. X:# de alumnos con enfermedad E de 80.
$X\sim B(80,0.041)$, $P[X\leq2]=0.33$ (0.2358 sin corregir).

11. Un distribuidor recibe al mismo tiempo un lote grande de mercadería en cajas. El 60% proviene del proveedor A y el resto proviene del proveedor B. Por experiencias anteriores se sabe que el porcentaje de cajas que contiene a lo más una unidad defectuosa es 1% de A y 2% de B. Si del lote se escogen al azar 36 cajas, ¿Cuál es aproximadamente la probabilidad de que a lo más dos de éstas contengan a lo más una unidad defectuosa?

Rp. $p=0.01\times0.6+0.02\times0.4=0.014$. X: # de cajas que contienen una o más unidades defectuosas,
$X\sim B(36,p)$, $P[X\leq 2]=0.9977$ (0.9830 sin corregir)..

12.a) Un sistema está formado por 100 componentes que funcionan independientemente. La probabilidad de que cualquier componente falle durante el periodo de operación es 0.1. El sistema funciona si al menos 80 componentes funcionan. Calcular la probabilidad de que el sistema funcione.

b) Suponga que el sistema anterior está formado por n componentes, cada una con probabilidad de funcionamiento de 0.9. El sistema funciona si al menos el 80% de las componentes funcionan. Determine el valor de n de modo que el sistema funcione con probabilidad 0.9772.

Rp. Exito: Componente funciona. a) X: # éxitos de 100, $X\sim B(100, 0.9)$, $P[X\geq80]\cong0.9998$,
b) X: # éxitos de n, $X\sim B(n, 0.9)$, $P[X/n\geq 0.8]\cong0.9772$, $n=36$

13. Una compañía hotelera observa que el 12% de habitaciones reservadas no son cubiertas. La compañía decide aceptar reservas por un 10% más de las 450 habitaciones que dispone. Calcular aproximadamente el porcentaje de clientes con reservas que se quedarán sin habitación.

Rp. X :# habitaciones reservadas cubiertas de $n=495$ (450 + 10%(450), $X\sim B(495, 0.88)$,
$P[X>450]\cong0.0197$..

14. Cincuenta alumnos harán uso de las computadoras del laboratorio de computo. Se ha estimado que cada alumno usará una computadora el tiempo promedio de 36 minutos por hora solicitada, ¿cuántas computadoras debería tener el

laboratorio de tal manera que en cualquier instante el número de computadoras sea insuficiente con probabilidad 0.0217?.

Rp. Sea k el # de computadoras. X: # de alumnos de n=50 que usarán las computadoras, X~B(50,0.6), p=0.6 (36 minutos de 60). El número de computadoras es insuficiente si X>k, $P[X>k]=0.0217$, $k\cong37$.

15. Un artículo eléctrico tiene una duración promedio de 800 horas. También se sabe que dicha duración es una variable aleatoria que tiene distribución exponencial. El costo de fabricar el artículo es de S/.200. El fabricante vende el artículo a S/.500 pero garantiza la devolución total del dinero si dura menos de 1000 horas.

a) ¿Cuál es la utilidad esperada del artículo?

b) De 200 artículos, ¿cuál es la probabilidad de que más del 65% sean devueltos con garantía?,

Rp. a) $E(U)=-200(1-e^{-10/8})+300e^{-10/8}=-56.75$, b) X~B(200, p). $P[X>130]\cong0.9767$.

16. El costo de la producción de cierto artículo tiene una media de $8 y una desviación estándar de 1$. Si se adquieren 36 de tales artículos para su comercialización, hallar el valor de la venta K de cada uno de ellos de manera que al vender los 36 se quiere ganar al menos $98.13 con probabilidad 0.95.

Rp. K=11$.

17. Un dado se lanza 300 veces. Calcular la probabilidad de que el número de veces que se obtiene el 4 difiera del número de veces que se espera el 4 en más de 12.

Rp. $P[|X-E(X)|>12]=0.0524$ (0.0628 sin corregir).

18. Una máquina es dada de baja para ser reparada, si una muestra aleatoria de 100 artículos de su producción da por lo menos 2% de artículos defectuosos. Calcular la probabilidad de que la máquina sea dada de baja si realmente produce el 2% de artículos defectuosos.

Rp. $P[X\geq2/p=0.02]=0.6406$ (0.5000 sin corregir).

19. Un fabricante de cierta medicina sostiene que ésta cura cierta enfermedad en el 90% de los casos. Para verificar tal afirmación se utiliza el medicamento en una muestra de 100 pacientes con esa enfermedad y se decide aceptar la afirmación si se curan 80 o más, en caso contrario se rechaza.

a) ¿Cuál es la probabilidad de aceptar la afirmación cuando la probabilidad de curación sea 0.80?.

b) ¿Cuál es la probabilidad de rechazar la afirmación cuando la probabilidad de curación sea 0.90?.

Rp. a) 0.5498 (sin corrector: 0.5000), b) 0.0002 (sin corrector 0.0001)

20. Para controlar la calidad de un proceso de producción de cierto tipo de objetos, cada vez se selecciona una muestra aleatoria de tamaño 40 uno por uno sin reposición. Si el número de objetos defectuosos en la muestra es al menos k se

detiene el proceso, en caso contrario se continua. Calcular el valor de k para que con probabilidad (corregida) de 0.0351 no se continúe con el proceso cuando la producción total contenga 5% de objetos defectuosos. Rp. $k = 5$.

21. Un transbordador transporta 300 pasajeros. Se sabe que el peso de la población de pasajeros tiene una media de 63 kgs. y una varianza de 135 kgs^2 y que el 30% de toda la población de pasajeros tiene pesos que superan los 90 kgs.

a) Si el reglamento de seguridad establece que el peso total de los pasajeros del transbordador no debe exceder los 19000 kgs en más del 5% de las veces. ¿Cumple el transbordador las reglamentaciones de seguridad?.

b) Si la población consiste de 2000 pasajeros, ¿qué probabilidad hay de que menos del 25% de los 300 pasajeros tengan un peso que supere los 90 kilogramos?

Rp. a) $Y_{300} \sim N(18900, (201.246)^2)$, $P[Y_{300}>19000]=P[Z>(19000-18900)/201.246]= P[Z>0.50]=0.309$ No cumple. b) $X\sim H(2000, 600,300)$, $P[X\leq 74]=P[Z\leq(74.5-90)/7.3196]=0.0170$

22. Suponga que el 20% de las llamadas telefónicas de un vendedor resultan en ventas. Halle la probabilidad de que sean necesarias efectuar al menos 140 llamadas para que resulten 40 ventas.

Rp. X: # llamadas hasta tener 40 ventas, $E(X)=r/p=40/(0.2)=200$. $V(X)=r(q/p^2)=40[0.8/(0.2)^2]=800$ $P[X\geq 140]= 0.9838$.

23. En un río de la selva donde existen muchos tipos de peces, se sabe que el 40% de ellos son de la especie E. Si los peces se pescan uno por uno, ¿cuál es la probabilidad de que sean necesarios pescar más de 60 peces para obtener 36 de la especie E?

Rp. X: # intentos hasta obtener 36 de especie E, $E(X)=36/(0.4)=90$, $V(X)=36[0.6/(0.4)^2]=135$. $P[X>60]=0.9945$.

24. El monto de las compras diarias de los clientes de la tienda P&C que adquirieron una tarjeta de crédito especial tiene una media de 250 soles y una desviación estándar de 60 soles. Además el 12.5% de todos los clientes compra por más de $300.

a) El gerente de ventas de esta tienda afirma que el total de compras diarias de los primeros 36 clientes con tarjeta de crédito especial supera los 10,000 soles en más del 5% de las veces. ¿Es correcta la afirmación del Gerente?

b) Si la población consiste de 400 clientes de la tienda que tienen la tarjeta de crédito especial ¿Qué probabilidad hay de que al menos el 25% de 36 de estos clientes escogidos al azar compren por más de 300 soles?

Rp. a) $Y_{36}\cong N(9000,(360)^2)$ $P[Y_{36}>10000]=P[Z>2.78]=0.0027$.
b) $H(400,50,36)$, c) $Z=(X-4.5)/1.8952$, $P[X\geq 9]=P[Z\geq 2.37]=0.0089$

25. El promedio del número de accidentes de trabajo en una fábrica es 1 cada semana.

a) Si acaba de ocurrir un accidente, ¿cuál es la probabilidad de que transcurra hasta 2 semanas antes de que ocurra otro accidente?

b) Calcular la probabilidad de que se produzcan más de 45 accidentes durante 36 semanas.

Rp. a) X: Poisson, $\lambda=1$, b) T:Exponencial con $\beta=\lambda=1$, $P[X\leq 2]=1-e^{-2}$.
b)$P[X>45]=P[Z\geq(45.5-36\times 1)/6]=0.0571$

26. Una empresa produce láminas de acero de 4 metros. El número de defectos que se encuentra al desenrollar las láminas es una variable aleatoria de Poisson que tiene una media de 2 defectos por lámina. ¿Cuál la probabilidad de que, se encuentren a lo más 50 defectos al desenrollar 32 rollos de lámina?.

Rp. $\lambda=2$, $P[X\leq 50]=P[Z\leq(50.5-32\times 2)/8]=0.9545$

27. La vida de un chip es una variable aleatoria que tiene distribución exponencial con parámetro $1/\theta$.

a) Calcule el porcentaje de chips que funcionan al menos el tiempo promedio.
b) Determine la probabilidad aproximada de que el tiempo de vida útil total de 36 chips sea superior a 48θ

Rp. a) $E(X)=\theta$, $P[X>\theta]=e^{-1}$. b) $P[Y_{36}>48\theta]=P[Z>(48\theta-36\theta)/6\theta]==P[Z>2]=0.0228$

28. Una máquina llena vasos de refrescos de 8 onzas. El número de onzas por vaso tiene una distribución normal con una media de 7.8 onzas y con una desviación estándar de 0.8 onzas. ¿Qué probabilidad hay de que al llenar 100 vasos de 8 onzas el líquido total derramado sea mayor de 6 onzas?

Rp. $X\sim N(7.8, (0.8)^2)$, $Y_{100}\sim N(780, 8^2)$, b) $P[Y_{100}>806]=P[Z>(806-780)/8]=P[Z>3.25]=0.0006$.

29. El tiempo de funcionamiento de un determinado tipo de batería es una variable aleatoria con distribución gamma de media 5 semanas y desviación estándar 1.5 semanas. En cuanto la batería deja de funcionar se reemplaza por otra nueva con las mismas características.

a) Justificando debidamente, determine la probabilidad aproximada de que en cinco años (52 semanas por año) no hayan sido suficiente 50 baterías.
b) ¿Con cuántas baterías se asegura que el tiempo total de funcionamiento es al menos 162 semanas con probabilidad 0.9772?

Rp. a) T_{50}=func. 50 baterías , $T_{50}\cong N(50\times 5, 50\times(1.5)^2)$, $P[T_{50}<52\times 5]=P[Z<0.9428]=0.8264$.
b) n? / $P[T_n\geq 162]=0.9772$, $n=36$

30. El departamento de créditos de la tienda comercial "TOP y TOP" sabe que el 30% de sus ventas son pagadas con dinero en efectivo, 30% con cheque y el 40% al crédito. La probabilidad de que una venta sea por más de \$50 es igual a 0.2 si ésta es en efectivo, es igual a 0.9 si ésta es con cheque y es igual a 0.6 si ésta es al crédito. Un investigador escoge una muestra aleatoria de 256 clientes que acaban de comprar en esa tienda.. Halle la probabilidad de que menos de 126 clientes de la muestra hayan comprado por más de \$50

Rp. $p=0.3\times 0.2+0.3\times 0.9+0.4\times 0.6=0.57$, $X\sim B(250, p)$, $P[X\leq 125]=P[Z<-2.58]=0.0049$

Capítulo 8

DISTRIBUCIONES MUESTRALES

8.1 Muestreo aleatorio

8.1.1 Población y parámetros

Definición. Se denomina **población** o universo a la totalidad de personas u objetos que tienen una o más características medibles o contables de naturaleza cualitativa o cuantitativa.

La característica medible o contable es una *variable estadística* cuyo valor, numérico o no numérico, es una *observación*.

Si la variable estadística a estudiar es una sola, cada elemento de la población puede asociarse con una observación. En este sentido, se denomina población al conjunto de valores posibles de la variable.

Si los elementos de la población se definen en forma aleatoria, entonces la variable estadística cuantitativa es una variable aleatoria cuyos valores constituyen la población. En este caso, la distribución de la población es la distribución de la variable aleatoria, por lo tanto, la media y la varianza de la variable aleatoria, vienen a ser la media y la varianza de la población.

Si la variable aleatoria X tiene distribución $f(x)$, se puede referir a la población $f(x)$, Por ejemplo, si X está normalmente distribuida se dice que la población está normalmente distribuida o que se tiene una población normal.

Por el número de observaciones la población puede ser *finita* de tamaño N, o *infinita*. Algunas poblaciones finitas son tan grandes que en teoría son asumidas como poblaciones infinitas.

Definición. Se denominan **parámetros** a las medidas descriptivas que caracterizan a la distribución de la población. Entre otros, los parámetros poblacionales son.

Media : μ
Proporción : π o p
Varianza : σ^2
Desviación estándar: σ

En diversas aplicaciones estadísticas al estudiar una población, la variable aleatoria que la define puede tener distribución conocida o no. La distribución de la población es conocida, si se conocen sus parámetros y su forma, es decir si se conoce su distribución de probabilidad .

Si la distribución de la población es desconocida, podemos estar interesados en:

* **Estimar sus parámetros**, si se conoce su distribución, y
* **Probar determinada suposición** acerca de un valor determinado del parámetro, o probar la suposición acerca del tipo de distribución de probabilidades de la población.

8.1.2 Muestra aleatoria

En vez de examinar la población entera, lo cual puede resultar físicamente imposible o no práctica, puede examinarse una **muestra** de la población con el propósito de **inferir** los resultados encontrados.

Una muestra es un subconjunto de la población.

El proceso de selección de una muestra de *n* elementos de la población se llama **muestreo**. Las ventajas y las razones para el muestreo son diversas, las mismas que no explicaremos en este texto.

El proceso que consiste en inferir resultados a la población a partir de la muestra se denomina **inferencia estadística.** La confiabilidad de las conclusiones extraídas concernientes a una población dependen de si la muestra se ha escogido apropiadamente de manera que represente bien a la población.

Una técnica para obtener muestras representativas de la población es el **muestreo aleatorio**. Se llama muestreo aleatorio a todo proceso que asegure en cualquier momento del mismo igual probabilidad de ser incluidos en la muestra a todos los elementos que pertenezcan a la población en dicho momento.

A las muestras aleatorias se les denomina también muestras probabilísticas

Las muestras aleatorias son de 4 tipos: Al azar simple, al azar sistemático, estratificado y por grupos (o conglomerados).

Muestra al azar simple

Es aquella en la que los elementos se escogen del total de la población en forma individual con una oportunidad igual e independiente .Por lo general se utiliza una tabla de números aleatorios.

Si la población es infinita el muestreo aleatorio ocurre cuando la extracción de los elementos de la muestra se hace con o sin reemplazo. Si la población es finita

de tamaño *N*, el muestreo aleatorio ocurre también si la extracción es con o sin reemplazo. Con reemplazo, la probabilidad de cada elemento de ser extraído es 1/N. Si es sin reemplazo, la probabilidad de cada elemento de ser elegido es 1/N en la primera extracción, es de 1/(N−1) en la segunda extracción, es 1/(N−2) en la tercera extracción, etc.

Por **ejemplo**, seleccionar una muestra al azar simple es similar a la que se realiza en la extracción aleatoria de números en una lotería.

Muestra al azar sistemática

Una muestra aleatoria sistemática es aquella en que sus elementos se eligen de la población a intervalos uniformes a partir de un listado ordenado. El *k*-ésimo elemento de la muestra es $k=N/n$, donde n es el tamaño de la muestra y N el tamaño de la población.

Por **ejemplo**, al elegir una muestra sistemática de 100 alumnos de EE.GG.CC que tiene 3000 alumnos, k =3000/100 =30. El primero se elige en forma aleatoria de los 30 primeros de la lista y los demás sistemáticamente cada 30 alumnos de la lista.

Muestreo aleatorio estratificado

Primero se clasifican a los elementos de la población en subgrupos separados de acuerdo con una o más características importantes (estratos). Después se obtiene por separado una muestra aleatoria simple o sistemática en cada estrato. El tamaño de cada submuestra debe ser proporcional al tamaño del estrato para asegurar representatividad.

Por **ejemplo**, para obtener una muestra aleatoria de 600 electores de una población de 600,000 electores de los cuales 300,0000 son de clase baja, 200,000 de clase media y 100,000 de clase alta. Se deben elegir al azar 300 de clase baja, 200 de clase media y 100 de clase alta.

Muestreo aleatorio agrupado

Denominado también por conglomerados. Los elementos de la población se dividen en forma natural en subgrupos. Luego se eligen al azar los subgrupos que forman la muestra.

Por **ejemplo**, al estudiar los pensiones que se pagan en los colegios particulares donde no es posible tener una lista de todas las pensiones, pero puede obtenerse una lista de los colegios particulares (grupos). Entonces, con esta lista puede obtener una muestra aleatoria de colegios y así obtener los pensiones que se pagan en estos colegios.

El **muestreo aleatorio simple**, es pues el proceso de selección de una muestra por el cual cada uno de los elementos de la población tienen una oportunidad igual e independiente de ser incluidos en la muestra. En el muestreo aleatorio simple cada

variable aleatoria X_i cuyo valor es x_i, tiene la misma distribución de la población de la cual se obtiene. Por ejemplo, supongamos que una población consiste de 8 fichas, dos con el número 2, cuatro con el número 5, y dos con el número 7. Si se extrae una ficha al azar, la ficha puede tomar cualquiera de los tres valores: 2 con probabilidad 0.25, 5 con probabilidad 0.50, y 7 con probabilidad 0.25, que viene a ser la misma distribución de la población.

Luego, diremos que los valores $x_1, x_2, ... x_n$ tomados respectivamente por las variables aleatorias $X_1, X_2, ..., X_n$, constituyen una muestra aleatoria simple de tamaño n de una población $f(x)$ de la variable aleatoria X, si estas variables aleatorias están distribuidas en forma idéntica a la distribución de la población y son independientes. Llamaremos también muestra aleatoria simple a este conjunto de variables aleatorias. Formalmente definimos una muestra aleatoria simple o brevemente muestra aleatoria de la forma siguiente:

Definición. (muestra aleatoria simple). Dada una población $f(x)$ con media μ y varianza σ^2, se denomina muestra aleatoria de tamaño n de esa población, a un conjunto de n variables aleatorias $X_1, X_2, ..., X_n$ tales que:

1) Son independientes.
2) Cada una de ellas está distribuida en forma idéntica a $f(x)$.

La condición 1) implica que la **distribución de probabilidad conjunta** de la muestra aleatoria $X_1, X_2, ..., X_n$ es la expresión:

$$f(x_1, x_2, ..., x_n) = f(x_1) f(x_2) ... f(x_n)$$

La condición 2) significa que:

a) Cada variable aleatoria X_i tiene la misma media y varianza de la distribución de X, es decir: $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$.

b) La distribución de probabilidad de cada variable aleatoria X_i es la misma distribución de probabilidades de X, esto es, $f(x_i) = f(x)$.

NOTA. El proceso de obtener este tipo de muestra requiere población infinita o bien población finita pero con reposición de elementos.

EJEMPLO 8.1.

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n de una población normal $N(\mu, \sigma^2)$.

a) Escribir la función de densidad de probabilidad conjunta de la muestra.

b) Si $n = 6$, $\mu = 20$, y $\sigma^2 = 25$, calcular la probabilidad de que:

b1) $X_1 + X_3 + X_4 - X_6$ sea mayor que 52.

b2) al menos una de las X_i sea menor que 29.8.

SOLUCION.

a) La función de densidad conjunta de la muestra aleatoria es

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n) = [f(x_i)]^n$$

$$f(x_1, x_2, \ldots x_n) = \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right]^n = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

b1) La media y la varianza de la variable aleatoria $Y = X_1 + X_3 + X_4 - X_6$ están dadas respectivamente por:

$$E(Y) = E(X_1) + E(X_3) + E(X_4) - E(X_6) = 20 + 20 + 20 - 20 = 40.$$

$$V(Y) = V(X_1) + V(X_3) + V(X_4) + V(X_6) = 25 + 25 + 25 + 25 = 100.$$

Por la propiedad reproductiva de la normal la variable aleatoria Y tiene distribución normal $N(40,100)$, luego, la variable aleatoria estándar:

$$Z = \frac{Y - \mu_Y}{\sigma_Y} = \frac{Y - 40}{10}, \quad \text{tiene distribución} \quad N(0,1), \quad \text{y}$$

$$P[Y > 52] = P\left[\frac{Y - 40}{10} > \frac{52 - 40}{10}\right] = P[Z > 1.2] = 0.1151.$$

b2) Sea ahora la variable aleatoria

$$Y_i = \begin{cases} 1, & \text{si } X_i < 29.8 \\ 0, & \text{si } X_i \geq 29.8 \end{cases},$$

Entonces, Y_i es Bernoullí $B(1, p)$, donde $p = P[Y_i = 1]$, la probabilidad del éxito es igual a:

$$p = P[X_i < 29.8] = P\left[\frac{X_i - 20}{5} < \frac{29.8 - 20}{5}\right] = P[Z < 1.96] = 0.975.$$

En consecuencia, la variable aleatoria: $Y=\sum_{i=1}^{6} Y_i$ es binomial $B(6, p)$, esto es,

$$P[Y=y]=C_y^6 p^y(1-p)^{6-y}, \quad y=0,1,2,3,4,5,6,$$

Por tanto, la probabilidad de que al menos un X_i, sea menor que 29.8 es:

$$P[Y \geq 1]=1-P[Y=0]=1-(0.025)^6=1-0.000=1.000.$$

8.1.3 Estadísticas

Definición. Se denomina **estadística** a cualquier función de las variables aleatorias que constituyen la muestra.

Una estadística es una variable aleatoria $Y=H(X_1, X_2, ..., X_n)$, cuyo valor es el número real $y=H(x_1, x_2, ... x_n)$. El término estadística se usa para referirse tanto a la función de la muestra, como al valor de esta función.

En general para cada parámetro poblacional hay una estadística correspondiente a calcularse a partir de la muestra. Algunas estadísticas importantes y sus valores calculados a partir de una muestra aleatoria son:

a) La media muestral $\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$, con valor: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$.

b) La varianza muestral $S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$, con valor: $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

c) La desviación estándar muestral $S=\sqrt{S^2}$.

d) La proporción muestral (porcentaje de éxitos en la muestra) $\hat{P}$ o $\bar{P}=\frac{1}{n}\sum_{i=1}^{n} X_i$,

donde $X_i \sim B(1, p)$ (el parámetro p es el porcentaje de éxitos en la población). También,

$$\bar{P}=\frac{X}{n}, \quad \text{donde } X \sim B(n, p)$$

El valor de $\bar{P}$ (o $\hat{P}$), calculada a partir de una muestra, se denota por $\bar{p}$ (o $\hat{p}$).

8.2 Distribuciones muestrales

Definición. Se denomina **distribución muestral** de una estadística a su distribución de probabilidad.

Por ejemplo, a la distribución de probabilidad de la estadística media $\overline{X}$, se le denomina distribución muestral de la media.

8.2 .1 Distribución muestral de la media $\overline{X}$

TEOREMA 8.1. Sea $X_1, X_2, ..., X_n$, una muestra aleatoria de tamaño n escogida de una población $f(x)$ con media μ y con varianza σ^2. Si $\overline{X}$ es la media muestral , entonces,

a) $E(\overline{X}) = \mu$ b) $Var(\overline{X}) = \dfrac{\sigma^2}{n}$

c) Para n suficientemente grande, la variable aleatoria,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

tiene distribución aproximadamente normal $N(0,1)$.

PRUEBA.

Por la definición de muestra aleatoria, las variables aleatorias $X_1, X_2, ..., X_n$, son independientes e idénticamente distribuidas como $f(x)$ con $E(X_i) = \mu$, y con $Var(X_i) = \sigma^2$. Entonces,

a) $E(\overline{X}) = E\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i\right) = \dfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \dfrac{1}{n} n\mu = \mu.$

b) $Var(\overline{X}) = V\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i\right) = \dfrac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \dfrac{1}{n^2} n\sigma^2 = \dfrac{\sigma^2}{n}.$

c) Se deduce del teorema del límite central escribiendo

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}.$$

NOTAS.

1. La aproximación de $\overline{X}$ a la normal $N(\mu,\sigma^2/n)$ es buena si $n \geq 30$, sin importar si la población es discreta o continua.

2. Si la muestra aleatoria es escogida de una población normal $N(\mu,\sigma^2)$, entonces, la distribución de $\overline{X}$ es exactamente normal $N(\mu,\sigma^2/n)$, para cualquier tamaño de muestra, $n \geq 2$.

3. La varianza de la media: $Var(\overline{X}) = \dfrac{\sigma^2}{n}$ es válida, si el ***muestreo es con o sin reemplazo*** en una población infinita, o es con reemplazo en una población finita de tamaño *N*.
Si el muestreo es sin reemplazo en una población finita de tamaño *N*, entonces, la varianza de la distribución de $\overline{X}$ es:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

El coeficiente $\dfrac{N-n}{N-1}$ se denomina **factor de corrección para población finita**. *Observar* que cuando $N \to +\infty$ el factor de corrección tiende a uno.

4. La desviación estándar de una estadística es conocida como **error estándar**.

EJEMPLO 8.2.

Suponga que una población finita *X* consiste de los valores:

3, 4, 7, 9, 12.

a) Calcular la media y la varianza de la población.
b) Determinar la distribución muestral de la media de las muestras de tamaño dos escogidas con reposición.
c) Determinar la distribución muestral de la media de las muestras de tamaño dos escogidas sin reposición.
d) Si se extraen muestras al azar de tamaño 36 con reposición, ¿cuál es la probabilidad de que la media muestral esté entre los valores 5 y 8?

SOLUCION.

a) La distribución de probabilidad de esta población finita de tamaño *N*=5, es la distribución uniforme siguiente:

x_i	3	4	7	9	12
$f(x_i)=P[X=x_i]$	1/5	1/5	1/5	1/5	1/5

La media y la varianza de la población son respectivamente:

$$\mu=\sum_{i=1}^{N}x_i f(x_i)=\sum_{i=1}^{N}\frac{x_i}{N}=\frac{\sum_{i=1}^{N}x_i}{N}=\frac{3+4+7+9+12}{5}=7.$$

$$\sigma^2=\sum_{i=1}^{N}x_i^2 f(x_i)-\mu^2=\frac{\sum_{i=1}^{N}x_i^2}{N}-\mu^2=\frac{3^2+4^2+7^2+9^2+12^2}{5}-7^2=10.8.$$

b) Se pueden extraer $5\times 5=25$ muestras de tamaño dos con reposición. Las muestras y sus medias son las siguientes.

Muestras					Medias de las muestras				
3 ,3	3, 4	3, 7	3, 9	3, 12	3	3.5	5	6	7.5
4 ,3	4, 4	4, 7	4, 9	4, 12	3.5	4	5.5	6.5	8
7 ,3	7 ,4	7 ,7	7 ,9	7 ,12	5	5.5	7	8	9.5
9 ,3	9, 4	9, 7	9, 9	9, 12	6	6.5	8	9	10.5
12 ,3	12, 4	12, 7	12, 9	12, 12	7.5	8	9.5	10.5	12

La distribución de probabilidades de las medias es:

$\bar{x}$	3	3.5	4	5	5.5	6	6.5	7	7.5	8	9	9.5	10.5	12
$f(\bar{x})$	1/25	2/25	1/25	2/25	2/25	2/25	2/25	1/25	2/25	4/25	1/25	2/25	2/25	1/25

Luego,

$$\mu_{\bar{X}}=E(\bar{X})=\sum f(\bar{x})\bar{x}=175/25=7.$$

$$\sigma^2_{\bar{X}}=Var(\bar{X})=\sum f(\bar{x})\bar{x}^2-\mu^2=\frac{1360}{25}-7^2=5.4.$$

Observar también que: $\sigma^2_{\bar{X}}=\dfrac{\sigma^2}{n}=\dfrac{10.8}{2}=5.4.$

c) Se pueden extraer $5\times 4=20$ muestras de tamaño 2 sin reposición.

Las muestras y sus medias respectivas son las siguientes:

Muestras				Medias de las muestras			
3, 4	3, 7	3, 9	3, 12	3.5	5	6	7.5
4, 3	4, 7	4, 9	4, 12	3.5	5.5	6.5	8
7, 3	7, 4	7, 9	7, 12	5	5.5	8	9.5
9, 3	9, 4	9, 7	9, 12	6	6.5	8	10.5
12, 3	12, 4	12, 7	12, 9	7.5	8	9.5	10.5

La distribución de probabilidades de la media es:

$\bar{x}$	3.5	5	5.5	6	6.5	7.5	8	9.5	10.5
$f(\bar{x})$	2/20	2/20	2/20	2/20	2/20	2/20	4/20	2/20	2/20

Luego,

$$\mu_{\bar{X}} = E(\bar{X}) = \sum f(\bar{x})\bar{x} = 140/20 = 7.$$

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = \sum f(\bar{x})\bar{x}^2 - \mu^2 = \frac{1061}{20} - 7^2 = 4.05.$$

Observar también que:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{10.8}{2}\left(\frac{5-2}{5-1}\right) = 4.05\ .$$

d) Sea $\bar{X}$ la media de las muestras de tamaño $n = 36$ con reposición. La estadística $\bar{X}$ tiene media y varianza respectivas:

$$\mu_{\bar{X}} \text{ o } E(\bar{X}) = \mu_X = 7, \quad \text{y} \quad \sigma^2_{\bar{X}} \text{ o } Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{10.8}{36} = 0.3$$

El error estándar de $\bar{X}$, es $\sigma_{\bar{X}} = \sqrt{0.3} = 0.55$

Entonces, la variable estándar,

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - 7}{0.55}$$

tiene distribución aproximadamente normal $N(0,1)$. Por tanto,

$$P[5 \le \bar{X} \le 8] = P\left[\frac{5-7}{0.55} \le \frac{\bar{X}-7}{0.55} \le \frac{8-7}{0.55}\right] = P[-3.64 \le Z \le 1.82] = 0.9655.$$

EJEMPLO 8.3.

El número de automóviles por familia en una ciudad es una variable aleatoria X cuya distribución de probabilidad es como sigue:

x	0	1	2	3	4
$f(x)$	4/12	4/12	2/12	1/12	1/12

si se escoge al azar una muestra de 49 familias, ¿cuál es la probabilidad de que la media muestral de autos por familia esté entre 1 y 2?.

SOLUCION.

La media y la varianza de X son respectivamente:

$$\mu_X = E(X) = \sum x_i f(x_i) = 0(\frac{4}{12}) + 1(\frac{4}{12}) + 2(\frac{2}{12}) + 3(\frac{1}{12}) + 4(\frac{1}{12}) = 1.25.$$

$$\sigma_X^2 = E(X^2) - (\mu_X)^2 = \sum x_i^2 f(x_i) - (\mu_X)^2 = 3.08 - (1.25)^2 = 1.52.$$

Sea $\overline{X}$ la media del número de automóviles en muestras de 49 familias. La estadística $\overline{X}$ tiene distribución aproximadamente normal con media y varianza respectivas:

$$\mu_{\overline{X}} \text{ o } E(\overline{X}) = \mu_X = 1.25$$

$$\sigma_{\overline{X}}^2 \text{ o } Var(\overline{X}) = \frac{\sigma^2}{n} = \frac{1.52}{49} = 0.031.$$

El error estándar de la media $\overline{X}$, es $\sigma_{\overline{X}} = \sqrt{0.031} = 0.176$

Entonces, la variable estándar,

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 1.25}{0.176}$$

tiene distribución aproximadamente normal $N(0,1)$. Por tanto,

$$P[1 \le \overline{X} \le 2] \cong P\left[\frac{1-1.25}{0.176} \le Z \le \frac{2-1.25}{0.176}\right] \cong P[-1.42 \le Z \le 4.26] = 0.9222.$$

EJEMPLO 8.4

Un auditor toma una muestra aleatoria de tamaño $n = 100$ de un conjunto de 500 cuentas por cobrar. El auditor sabe que las 500 cuentas por cobrar constituyen una población finita cuya desviación estándar es $\sigma = \$145$. ¿Cuál es la probabilidad de que la media muestral difiera de la media poblacional en mas de \$26?

SOLUCION.

Sea $\overline{X}$ la media de la muestra de tamaño $n = 100$ escogida de la población finita de $N = 500$ casos. Entonces, la variable aleatoria $\overline{X}$ tiene distribución aproximadamente normal con media $\mu_{\overline{X}} = \mu$ y error estándar:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{145}{\sqrt{100}}\sqrt{\frac{500-400}{500-1}} = 12.982$$

En consecuencia, la variable aleatoria estándar:

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{12.982}$$

tiene distribución normal $N(0,1)$. La probabilidad de que la media muestral difiera de la media poblacional en más de \$26 es:

$$P\left[\left|\overline{X} - \mu\right| > 26\right] = P\left[|Z| > \frac{26}{12.982}\right] = P\left[|Z| > 2.00\right] = 0.0456$$

EJEMPLO 8.5.

Una máquina empaqueta un determinado producto, en paquetes cuyo peso, en gramos, se distribuye normalmente con una desviación estándar de 20 gramos y con una media μ que debe ser bien regulada.

a) La media μ está bien regulada si sólo el 1% de los pesos de todos los paquetes que produce la máquina tienen pesos mayores a 546.6 gramos, ¿cuánto vale μ?.

b) Con la media bien regulada, se programa el siguiente control del peso del producto: Cada hora se escogen al azar 4 paquetes, si el promedio de los pesos no está entre 480 y 520 gramos, se para la máquina para mantenimiento. En caso contrario se continua con el proceso. ¿Cuál es la probabilidad de parar la máquina cuando realmente está bien regulada?

c) Si la máquina está bien regulada, ¿con qué tamaño de muestra se consigue que la media muestral sea a lo más 490.2 gramos con probabilidad igual a 0.025?.

SOLUCION.

Sea X el peso del producto empaquetado por la máquina, $X \sim N(\mu, 20^2)$.

a) Se debe calcular μ tal que $P[X > 546.6] = 0.01$

$$0.01 = P[X > 546.6] = P\left[Z > \frac{546.6 - \mu}{20}\right]$$

de donde resulta:

$$\frac{546.6-\mu}{20}=2.33, \quad \text{y} \quad \mu=500.$$

b) Sea $\overline{X}$ la media de la muestra de tamaño $n=4$. Por la propiedad reproductiva de la normal la variable aleatoria $\overline{X}$ tiene distribución exactamente normal con media $\mu_{\overline{X}}=500$ y error estándar:

$$\sigma_{\overline{X}}=\sigma/\sqrt{n}=20/\sqrt{4}=10.$$

Entonces, la variable aleatoria estándar:

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}=\frac{\overline{X}-500}{10}$$

tiene distribución normal $N(0,1)$. La probabilidad de no parar la máquina cuando realmente está bien regulada es:

$$P[480\leq \overline{X}\leq 520/\mu=500]=P\left[\frac{480-500}{10}\leq Z\leq\frac{520-500}{10}\right]$$

$$=P[-2\leq Z\leq 2]=0.9544.$$

Luego, la probabilidad de parar la máquina es: $1-0.9544=0.0456$.

c) Sea $\overline{X}_n$ la media de la muestra de tamaño n. Entonces, la distribución de $\overline{X}_n$ es normal con media igual a 500, y error estándar igual a $20/\sqrt{n}$. Se debe calcular n tal que $P[\overline{X}_n\leq 490.2]=0.0250$. Entonces,

$$0.0250=P[\overline{X}_n\leq 490.2]=P\left[Z\leq\frac{490.2-500}{20/\sqrt{n}}\right].$$

De donde resulta,

$$\frac{-9.8}{20}\sqrt{n}=-1.96$$

$$\sqrt{n}=4$$

$$n=16$$

8.2.2 Distribución muestral de la proporción

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n extraída de la población de Bernoullí $B(1, p)$, donde p es el porcentaje de éxitos en la población y sea

$$\overline{P} = \frac{X_1 + X_2 + ...X_n}{n} = \frac{X}{n}$$

la **proporción de éxitos en la muestra,** siendo, $X = X_1 + X_2 + \cdots + X_n$ una variable binomial $B(n, p)$, entonces,

a) $\mu_{\overline{P}} = E(\overline{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p.$

b) $\sigma^2_{\overline{P}} = V(\overline{P}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}[np(1-p)] = \frac{p(1-p)}{n}.$

c) Si n es suficientemente grande, entonces la variable aleatoria:

$$Z = \frac{\overline{P} - p}{\sqrt{p(1-p)/n}}.$$

tiene aproximadamente distribución $N(0,1)$.

NOTAS.

1. El **error estándar** de $\overline{P}$ es: $\sigma_{\overline{P}} = \sqrt{\frac{p(1-p)}{n}}$

2. Si la población es finita de tamaño N y el muestreo es sin reposición el error estándar (desviación estándar de la hipergeométrica) es:

$$\sigma_{\overline{P}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}.$$

Observar que si N es grande con respecto a n el factor de corrección $\frac{N-n}{N-1}$ se aproxima a la unidad.

3. Si n es suficientemente grande ($n \geq 30$)

$$P[\overline{P} \le c] \cong P\left[Z \le \frac{c - p}{\sigma_{\overline{P}}}\right].$$

Sin embargo aproximaciones satisfactorias se obtienen si se introduce el **factor de corrección por continuidad** $\frac{1}{2n}$. Luego,

$$P[\overline{P} \le c] \cong P\left[Z \le \frac{(c + 1/(2n)) - p}{\sigma_{\overline{P}}}\right].$$

4. Observar que las dos expresiones de Z

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\overline{P} - p}{\sqrt{p(1-p)}}.$$

donde X es binomial y $\overline{P}$ es el porcentaje de éxitos en la muestra, tienen distribución $N(0,1)$.

EJEMPLO 8.6.

En un proceso de producción el porcentaje de unidades defectuosas producidas es 4%. Para controlar el proceso, se revisan periódicamente los objetos producidos.

a) Calcular aproximadamente la probabilidad de que en una muestra aleatoria de 150 unidades revisadas se encuentren 6% defectuosos.

b) Si el proceso de producción se para al encontrar al menos 5% de unidades producidas al revisar muestras aleatorias de 100 objetos cada vez, ¿cuál es la probabilidad de que el proceso continúe si realmente produce 6% defectuosos del total de la producción?

SOLUCION

a) Sea $\overline{P} = X/150$, la proporción de artículos defectuosos en la muestra de 150 unidades, donde X, el número de unidades defectuosos en la muestra de 150 unidades es $B(150, 0.04)$.

Si se utiliza el modelo exacto para ejecutar el cálculo, se tiene:

$$P[\overline{P} = 0.06] = P\left[\frac{X}{150} = \frac{9}{150}\right] = P[X = 9] = C_9^{150}(0.04)^9(0.96)^{141}.$$

Si se utiliza el modelo aproximado para ejecutar el cálculo, aproximando a la normal por el teorema central del límite, se tienen 2 procesos:

Un método, es utilizar la aproximación binomial a la normal:

$$P[\overline{P}=0.06]=P[X=9]\cong P[8.5\le X\le 9.5]$$

$$\cong P\left[\frac{8.5-150(0.04)}{\sqrt{150(0,04)(0.96)}}\le\frac{X-np}{\sqrt{np(1-p)}}\le\frac{9.5-150(0.04)}{\sqrt{150(0,04)(0.96)}}\right]$$

$$\cong P\left[\frac{8.5-6}{2.4}\le Z\le\frac{9.5-6}{2.4}\right]=P[1.04\le Z\le 1.46]=0.0771$$

El otro método, es usar la propia distribución muestral de la proporción:

$$P[\overline{P}=0.06]=P\left[0.06-\frac{1}{2(150)}\le\overline{P}\le 0.06+\frac{1}{2(150)}\right]=P[0.0567\le\overline{P}\le 0.0633]$$

$$\cong P\left[\frac{0.0567-0.04}{0.016}\le Z\le\frac{0.0633-0.04}{0.016}\right]$$

$$\cong P[1.04\le Z\le 1.46]=0.0771.$$

b) Sea $\overline{P}=X/100$, la proporción de artículos defectuosos en la muestra de 100 unidades, donde X, el número de unidades defectuosos en la muestra de 100 unidades, es en este caso; $B(100, 0.06)$. Entonces,

$$P[\overline{P}<0.05/p=0.06]=1-P[\overline{P}\ge 0.05/p=0.06]=1-0.6628=0.3372$$

donde, se ha utilizado la aproximación normal de la distribución de $\overline{P}$:

$$P[\overline{P}\ge 0.05/p=0.06]\cong P\left[Z\ge\frac{0.05-0.06}{0.0237}\right]\cong P[Z\ge -0.42]=0.6628$$

EJEMPLO 8.7.

La probabilidad de que un paciente se recupere de una rara enfermedad es 0.4. ¿Cuál es la probabilidad de que en una muestra de 100 pacientes seleccionados de una población de 1,000 que sufren la enfermedad, más del 30% sobrevivan?.

SOLUCION.

Sea $\overline{P}=X/100$, la proporción de pacientes que se recuperan en la muestra de 100, donde X, es el número de pacientes que se recuperan en la muestra de 100.

Debido a que el muestreo es sin reposición, X tiene distribución de probabilidad hipergeométrica $H(1000, 400, 100)$

Para calcular $P[\overline{P} > 0.30]$ se pueden utilizar dos procesos. Uno de los métodos es la aproximación de la distribución hipergeométrica a la normal. En este caso:

$$Z = \frac{X - np}{\sqrt{np(1-p)\left((N-n)/(N-1)\right)}} = \frac{X - 100(0.4)}{\sqrt{100(0.4)(0.6)\left((1000-100)/(1000-1)\right)}} = \frac{X - 40}{4.6499}$$

tiene distribución aproximadamente normal $N(0,1)$.

El otro método es utilizar la distribución muestral de proporciones, en este caso:

$$Z = \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n} \times \frac{N-n}{N-1}}} = \frac{\overline{P} - 0.4}{\sqrt{\frac{(0.4)(0.6)}{100}\left(\frac{1000-100}{1000-1}\right)}} = \frac{\overline{P} - 0.4}{0.0465}$$

tiene también distribución aproximadamente normal $N(0,1)$.

Utilizando la distribución de la proporción muestral, se tiene:

$$P[\overline{P} > 0.30] = 1 - P[\overline{P} \leq 0.30] = 1 - 0.9842 = 0.0158$$

en donde,

$$P[\overline{P} \leq 0.3] \cong P\left[Z \leq \frac{0.3 - 0.4}{0.0465}\right] = P[Z \leq -2.15] = 0.0158$$

8.2.3 Distribución muestral de $\frac{n\hat{S}^2}{\sigma^2}$

TEOREMA 8.2. (Distribución muestral de la varianza)

Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria escogida de una **distribución normal** $N(\mu, \sigma^2)$, y si,

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$$

es la varianza muestral, entonces,

a) $E\left(S^2\right) = \frac{n-1}{n}\sigma^2$

b) $\dfrac{nS^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$ tiene distribución $\chi^2(n-1)$.

PRUEBA.

a) Probaremos primero que:

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n\left(\overline{X} - \mu\right)^2$$

En efecto,

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}\left(X_i - \mu + \mu - \overline{X}\right)^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 + 2(\mu - \overline{X})\sum_{i=1}^{n}(X_i - \mu) + n\left(\mu - \overline{X}\right)^2$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\mu - \overline{X})^2 + n(\mu - \overline{X})^2$$

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n\left(\overline{X} - \mu\right)^2.$$

Luego,

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\right) = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu)^2] - E[(\overline{X} - \mu)^2]$$

$$E(S^2) = \frac{1}{n}\left(n\sigma^2\right) - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

b) Probaremos primero que:

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - \mu\right)^2$$

En efecto,

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X} + \overline{X} - \mu\right)^2$$

$$= \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}\left(X_i - \bar{X}\right) + n\left(\bar{X} - \mu\right)^2$$

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - \mu\right)^2, \text{ ya que } \; 2(\bar{X} - \mu)\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$$

Dividiendo por σ^2 la identidad probada, resulta,

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} + n\frac{(\bar{X} - \mu)^2}{\sigma^2}$$

Por otra parte, se sabe que,

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \text{ tiene distribución } \chi^2(n) \text{ y}$$

$$n\frac{(\bar{X} - \mu)^2}{\sigma^2} = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \text{ tiene distribución } \chi^2(1)$$

Utilizando métodos avanzados mas allá de este libro, puede demostrarse que estas dos últimas variables son independientes. Luego, por la propiedad reproductiva de la distribución chi-cuadrado resulta que:

$$\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} \text{ tiene distribución } \chi^2(n-1)$$

NOTAS.

1) Si definimos la varianza, $\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$, entonces,

$$E\left(\hat{S}^2\right) = E\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}E\left(S^2\right) = \left(\frac{n}{n-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = \sigma^2.$$

2) En $\hat{S}^2 = \frac{n}{n-1}S^2$, se tiene que $\frac{n}{n-1} \to 1$, cuando $n \to +\infty$,

3) $\frac{(n-1)\hat{S}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$ tiene distribución $\chi^2(n-1)$.

4) Se verifica que: $\hat{S}^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2}{n-1}$

EJEMPLO 8.8.

Si S^2 es la varianza de una muestra aleatoria de tamaño $n = 15$ escogida de una población normal con media μ y varianza σ^2, calcular:

a) $P[0.3107 \le S^2/\sigma^2 \le 1.9427]$,

b) $P[0.3329 \le \hat{S}^2/\sigma^2 \le 2.0814]$

SOLUCION.

a) Con $n = 15$ la variable aleatoria $15S^2/\sigma^2$ tiene distribución chi-cuadrado con 14 grados de libertad, entonces,

$$P[0.3107 \le S^2/\sigma^2 \le 1.9427] = P[(15)(0.3107) \le 15S^2/\sigma^2 \le (15)(1.9427)]$$

$$= P[4.66 \le \chi^2(14) \le 29.14] = 0.99 - 0.01 = 0.98.$$

b) Con $n = 15$ la variable aleatoria $14\hat{S}^2/\sigma^2$ tiene distribución chi-cuadrado con 14 grados de libertad,

entonces,

$$P[0.3329 \le \hat{S}^2/\sigma^2 \le 2.0814] = P[(14)(0.3329) \le 14\hat{S}^2/\sigma^2 \le (14)(2.0814)]$$

$$= P[4.66 \le \chi^2(14) \le 29.14] = 0.99 - 0.01 = 0.98.$$

8.3 OTRAS DISTRIBUCIONES MUESTRALES

8.3.1 Distribución de $\frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}$

TEOREMA 8.3. (Distribución de $\overline{X}$ cuando σ^2 se desconoce)

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n escogida **de una distribución normal** $N(\mu, \sigma^2)$, donde la varianza poblacional σ^2 es desconocida. Entonces, la variable aleatoria:

$$T = \frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}.$$

tiene distribución t-student con $n-1$ grados de libertad, o $T \sim t(n-1)$

En efecto, se ha verificado que:

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad \text{y} \quad V = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1),$$

además las variables Z y V son independientes.

Entonces, la variable aleatoria:

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}},$$

tiene distribución t-student con $n-1$ grados de libertad.

NOTA. Observar que

$$V = \frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1)$$

Entonces, la variable aleatoria:

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n-1}},$$

tiene también distribución t-student con $n-1$ grados de libertad.

EJEMPLO 8.9.

Si $\overline{X}$ es la media y $\hat{S}^2$ es la varianza de una muestra aleatoria de tamaño $n=9$ seleccionada de una población normal con media $\mu = 90$, calcular

$$P[0.2353 \leq \frac{\overline{X}-90}{\hat{S}} \leq 1.1183].$$

SOLUCION.

En este caso, la variable aleatoria:

$$T = \frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}} = \frac{\overline{X}-90}{\hat{S}/\sqrt{9}} = \frac{3(\overline{X}-90)}{\hat{S}}$$

se distribuye según t-student con 8 grados de libertad, esto es, $T \sim t(8)$, entonces,

$$P[0.2353 \leq \frac{\overline{X}-90}{\hat{S}} \leq 1.1183] = P\left[(3)(0.2353) \leq \frac{3(\overline{X}-90)}{\hat{S}} \leq (3)(1.1183)\right]$$

$$= P\left[0.706 \leq t(8) \leq 3.355\right]$$

$$= 0.995 - 0.75 = 0.245$$

8.3.2 Distribución muestral de la diferencia de dos medias con varianzas poblacionales conocidas

TEOREMA 8.4. Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas de dos poblaciones con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 respectivamente, supuestas conocidas, entonces, la variable aleatoria $\overline{X}_1 - \overline{X}_2$ tiene las siguientes propiedades:

a) $\mu_{\overline{X}_1-\overline{X}_2} = E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2.$

b) $\sigma^2_{\overline{X}_1-\overline{X}_2} = V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}.$

c) Para n_1 y n_2 suficientemente grandes , la variable aleatoria:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tiene aproximadamente distribución normal $N(0,1)$ (Verificar!).

NOTA.

La aproximación de Z a la normal es muy buena si $n_1 \geq 30$ y $n_2 \geq 30$ sin importar si las poblaciones son discretas o continuas y sin importar sus formas.

Pero, si las dos poblaciones son normales, entonces, la media $\overline{X}_1$ es $N(\mu_1, \sigma_1^2/n_1)$ y $\overline{X}_2$ es $N(\mu_2, \sigma_2^2/n_2)$ para $n_1 \geq 2$ y $n_2 \geq 2$.

Por la propiedad reproductiva de la normal $\overline{X}_1 - \overline{X}_2$ es normal con media= $\mu_1 - \mu_2$ y varianza= $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

Luego, la distribución de la variable aleatoria Z:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

es normal $N(0,1)$ para cualquier valor de $n_1 \geq 2$ y $n_2 \geq 2$.

EJEMPLO 8.10.

Se han extraído dos muestras aleatorias del mismo tamaño n de dos máquinas que embolsan automáticamente un mismo producto cuya característica medible es el peso en gramos. Se sabe que los pesos de los productos de cada máquina se distribuyen normalmente con medias respectivas iguales a 120 gramos y con varianzas respectivas iguales a 18 gramos^2. Encontrar el valor de n de manera que la probabilidad de que las medias muestrales difieran en menos de 2 gramos sea 0.95.

SOLUCION.

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de los pesos de las muestras respectivas. Los pesos de las poblaciones se distribuyen normalmente con medias iguales y varianzas iguales. Entonces, la diferencia de las dos medias muestrales $\overline{X}_1 - \overline{X}_2$ tiene distribución normal con:

Media: $\mu_1 - \mu_2 = 0$, y varianza: $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{18+18}{n} = \frac{36}{n}$.

Luego, la variable

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{36/n}}$$ se distribuye como normal $N(0,1)$.

Se debe hallar el valor de n tal que $P\left[\left|\overline{X}_1 - \overline{X}_2\right| < 2\right] = 0.95$, entonces,

$$0.95 = P\left[\left|\overline{X}_1 - \overline{X}_2\right| < 2\right] = P\left[|Z| < \frac{2-0}{\sqrt{36/n}}\right] = P\left[-\frac{\sqrt{n}}{3} < Z < \frac{\sqrt{n}}{3}\right] = 2\Phi\left(\frac{\sqrt{n}}{3}\right) - 1,$$

de donde resulta,

$$2\Phi\left(\frac{\sqrt{n}}{3}\right) - 1 = 0.95, \quad \Phi\left(\frac{\sqrt{n}}{3}\right) = 0.975$$

$$\frac{\sqrt{n}}{3} = 1.96, \qquad \sqrt{n} = 5.88, \qquad n = 34.57 \cong 35.$$

8.3.3 Distribución muestral de la diferencia de dos medias con varianzas poblacionales desconocidas

Sea $\overline{X}_1$ la media de una muestra aleatoria de tamaño n_1 extraída de la **población normal** $N(\mu_1, \sigma_1^2)$, y sea $\overline{X}_2$ la media de otra muestra aleatoria de tamaño n_2 extraída de la **población normal** $N(\mu_2, \sigma_2^2)$, **independiente** de la anterior.

A) Varianzas poblacionales iguales: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

En este caso, la variable aleatoria $\overline{X}_1 - \overline{X}_2$ tiene distribución normal

$$N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

y la variable aleatoria estándar:

$$Z=\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

tiene distribución normal $N(0,1)$. Por otra parte,

$$\frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2}\sim\chi^2(n_1-1), \text{ y } \quad \frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$$

$$\text{y,} \quad V=\frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2}+\frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2}\sim\chi^2(n_1+n_2-2)$$

Por tanto, la variable aleatoria:

$$T=\frac{Z}{\sqrt{\frac{V}{n_1+n_2-2}}}=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{S}_C^2}{n_1}+\frac{\hat{S}_C^2}{n_2}}},$$

tiene distribución *t*-student con n_1+n_2-2 grados de libertad.

Donde, $\hat{S}_C^2$ **la varianza común**, tiene la expresión:

$$\hat{S}_C^2=\frac{(n_1-1)\hat{S}_1^2+(n_2-1)\hat{S}_2^2}{n_1+n_2-2}$$

B) Varianzas poblacionales diferentes $\sigma_1^2\neq\sigma_2^2$

En este caso la variable aleatoria:

$$T=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}},$$

tiene distribución *t*-student con *g* grados de libertad, donde,

$$g = \frac{\left[\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right]^2}{\frac{\left[\frac{\hat{S}_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{\hat{S}_2^2}{n_2}\right]^2}{n_2 - 1}}.$$

Si g no es un número entero se redondea al entero más cercano.

8.3.4 Distribución muestral de la diferencia de dos proporciones

Sean $X_1, X_2, ... X_{n_1}$ e $Y_1, Y_2, ..., Y_{n_2}$ dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones independientes de Bernoullí $B(1, p_1)$ y $B(1, p_2)$, donde p_1 y p_2 son las proporciones poblaciones de éxito respectivos. Sean las proporciones muestrales

$$\overline{P}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1} = \frac{X}{n_1} \quad \text{y} \quad \overline{P}_2 = \frac{\sum_{i=1}^{n_2} Y_i}{n_2} = \frac{Y}{n_2}$$

donde $X \sim B(n_1, p_1)$ y $Y \sim B(n_2, p_2)$

Entonces, la variable aleatoria $\overline{P}_1 - \overline{P}_2$ tiene una distribución de probabilidad cuyas propiedades son las siguientes:

a) $\mu_{\overline{P}_1 - \overline{P}_2} = E(\overline{P}_1 - \overline{P}_2) = E(\overline{P}_1) - E(\overline{P}_2) = p_1 - p_2.$

b) $\sigma^2_{\overline{P}_1 - \overline{P}_2} = V(\overline{P}_1 - \overline{P}_2) = V(\overline{P}_1) + V(\overline{P}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$

c) **Para n_1 y n_2 suficientemente grandes**, la variable aleatoria estándar:

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sigma_{\overline{P}_1 - \overline{P}_2}},$$

tiene distribución aproximadamente $N(0,1)$, donde el *error estándar* $\sigma_{\overline{P}_1-\overline{P}_2}$, es dado por:

$$\sigma_{\overline{P}_1-\overline{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

EJEMPLO 8.11.
Dos amigos A y B juegan "cara" o "sello" con una moneda. Suponga que en este juego, cada uno lanza la moneda 35 veces y que uno de ellos gana si obtiene 7 caras más que el otro. Calcular la probabilidad de que B gane el juego.

SOLUCION.

Sea X el número de caras que saca el jugador A en las 35 tiradas y sea Y el número de caras que saca el jugador B en las 35 tiradas, entonces, cada variable tiene distribución $B(35,0.5)$, donde $p = 0.5$ es la probabilidad de obtener cara en cada tirada de los jugadores.

Sean $\overline{P}_1 = X/35$ y $\overline{P}_2 = Y/35$ las proporciones de caras de los jugadores A y B respectivamente. El jugador B gana el juego si saca 7 caras más que A o si su proporción de caras es 7/35 = 0.2 más que A. Luego, la probabilidad de que B gane el juego es:

$P[\overline{P}_2 - \overline{P}_1 > 0.2] = 1 - P[\overline{P}_2 - \overline{P}_1 \leq 0.2] = 0.0475$, donde:

$$P[\overline{P}_2 - \overline{P}_1 \leq 0.2] \cong P\left[Z \leq \frac{0.2-(0.5-0.5)}{\sqrt{\frac{(0.5)(0.5)}{35} + \frac{(05)(0.5)}{35}}}\right] = P[Z \leq 1.67] = 0.9525$$

8.3.5 Distribución muestral de la razón $\hat{S}_1^2/\hat{S}_2^2$

TEOREMA 8.5. Si $\hat{S}_1^2$ y $\hat{S}_2^2$ son las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas de dos **poblaciones normales** $N(\mu_1,\sigma_1^2)$ y $N(\mu_2,\sigma_2^2)$ respectivas, entonces, la variable aleatoria

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene distribución F con grados de libertad n_1-1 y n_2-1, esto es, $F \sim F(n_1-1, n_2-1)$.

En efecto, la variable aleatoria $U=(n_1^{*}-1)\hat{S}_1^2/\sigma_1^2$ se distribuye como Chi^2 con n_1-1 grados de libertad y la variable aleatoria $V=(n_2-1)\hat{S}_2^2/\sigma_2^2$ se distribuye como Chi^2 con n_2-1 grados de libertad. Además son independientes. Luego la variable:

$$F=\frac{U/(n_1-1)}{V/(n_2-1)}=\frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2},$$

se distribuye según F con grados de libertad n_1-1 y n_2-1.

Observar que si $\sigma_1^2=\sigma_2^2$, entonces, $F=\hat{S}_1^2/\hat{S}_2^2$ se distribuye según F con grados de libertad n_1-1 y n_2-1, o $F \sim F(n_1-1, n_2-1)$

EJEMPLO 8.12.

De poblaciones $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$ independientes se extraen dos muestras de tamaños 13 y 21 respectivamente, hallar el valor de k tal que:

a) $P[\hat{S}_1^2 < k\hat{S}_2^2]=0.99$

b) $P[\hat{S}_1^2 / \hat{S}_2^2 < k]=0.05$

SOLUCION.

a) $P[\hat{S}_1^2 < k\hat{S}_2^2]=0.99$, implica $P[F<k]=0.99$, donde $F=\hat{S}_1^2/\hat{S}_2^2 \sim F(12, 20)$. Entonces, $k=3.23$.

b) $P[\hat{S}_1^2 / \hat{S}_2^2 < k]=0.05$, implica $P[F<k]=0.05$, donde $F=\hat{S}_1^2/\hat{S}_2^2 \sim F(12, 20)$. Entonces, $k=1/f_{0.95,20,12}=1/2.54=0.394$.

EJERCICIOS

Una media

1. De una población normal $N(6,6^2)$ se selecciona la muestra aleatoria: $X_1, X_2, \ldots, X_9$ de tamaño 9. Sea $\overline{X}$ la media de la muestra aleatoria
   a) Describa la distribución de probabilidades de $\overline{X}$.
   b) Determine el valor de c tal que $P[\overline{X} > c] = 0.985$.
   c) Si $Y = 3X - 5$, calcular $P[\overline{Y} > 28]$.

   Rp. a) $N(6, 36/9)$, b) $c = 1.66$, c) 0.0062.

2. Una población finita X consiste de los valores: 0, 2, 5, 8. Determine la distribución muestral de la media $\overline{X}$ para las muestras de tamaño dos escogidas de esta población:
   a) con sustitución, b) sin sustitución.

   Rp. a) 16 medias, $\mu_{\overline{X}} = 3.75$, $\sigma^2_{\overline{X}} = 4.59$, b) 12 medias, $\mu_{\overline{X}} = 3.75$, $\sigma^2_{\overline{X}} = 3.445$.

3. Suponga que los sueldos en cientos de dólares, en una región, es una variable aleatoria X cuya distribución de probabilidades es :

x	1	2	3	4	5
$f(x) = P[X = x]$	0.1	0.2	0.4	0.2	0.1

   si se toman al azar 30 sueldos de igual número de personas.
   a) halle la media y la varianza de la media muestral.
   b) calcule la probabilidad de que la media muestral esté entre 260 y 330 dólares.

   Rp. a) 3, 0.04, b) 0.9104.

4. La demanda diaria de un producto puede ser 0, 1, 2, 3, 4 con probabilidades respectivas 0.3, 0.3, 0.2, 0.1, 0.1.
   a) Describa la distribución de probabilidades aproximada de la demanda promedio de 36 días.
   b) Calcular la probabilidad de que la media de la demanda de 36 días esté entre 1 y 2 inclusive.

   Rp. a) Aproximadamente normal $N(1.4, 1.64/36)$, b) 0.9668.

5. De la historia sacada de los registros de la Universidad se ha determinado que las calificaciones del curso de MATE1 y de FILO1 se distribuyen normalmente con las medias respectivas 12 y 15 y con varianzas homogéneas igual a 4. ¿Cuál es la probabilidad de que el promedio las notas de un alumno en tales cursos esté, entre 13 y 16?.

   Rp. 0.5984

6. El gerente de ventas de una empresa cafetalera sabe que el consumo mensual de café por casa (en kilos) está normalmente distribuida con media desconocida μ y desviación estándar igual a 0.3. Si se registra el consumo de café durante un mes de 36 hogares escogidos al azar, ¿cuál es la probabilidad de que la media del consumo esté entre los valores $\mu - 0.1$ y $\mu + 0.1$?.

Rp. 0.9544

7. La distribución de las notas del examen final de Mat.I resultó ser normal $N(\mu, \sigma^2)$, con cuartiles 1 y 3 iguales a 6.99 y 11.01 respectivamente.

a) Determine la media y la varianza de la distribución de las notas

b) Halle el intervalo $[a, b]$ centrado en μ tal que $P[a \leq \overline{X} \leq b] = 0.9544$, donde $\overline{X}$ es la media de la muestra X_1, X_2, X_3, X_4 escogida de esa población.

Rp. a) $\mu = 9$, $\sigma = 3$. b) $a = 6$, $b = 12$.

8. La vida útil (en miles de horas) de una batería es una variable aleatoria X con función de densidad:

$$f(x) = \begin{cases} 2 - 2x & 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

Si $\overline{X}_{36}$ es la media de la muestra aleatoria $X_1, X_2, \ldots, X_{36}$ escogida de X, ¿con qué probabilidad $\overline{X}_{36}$ es mayor que 420 horas? Rp. 0.0136.

9. Sea $\overline{X}_{40}$ la media de la muestra aleatoria $X_1, X_2, \ldots, X_{40}$ de tamaño $n = 40$ escogida de una población X cuya distribución es geométrica con función de probabilidad:

$$f(x) = \frac{1}{5}\left(\frac{4}{5}\right)^{x-1}, x = 1, 2, \ldots$$

Hallar la probabilidad de que la media muestral difiera de la media poblacional en a lo más el 10% del valor de la varianza de la población

Rp. 0.9954.

10. La utilidad (en miles de soles) por la venta de cierto artículo, es una variable aleatoria con distribución normal. Se estima que en el 5% de las ventas la utilidad serían menos de 6.71, mientras que el 1% de las ventas serían mayores que 14.66. Si se realizan 16 operaciones de ventas, ¿cuál es la probabilidad de que el promedio de la utilidad por cada operación esté entre \$10,000 y \$11,000?.

Rp. 0.4772.

11. La vida útil de cierta marca de llantas radiales es una variable aleatoria X cuya distribución es normal con $\mu = 38{,}000$ Km. y $\sigma = 3{,}000$ Km.

a) Si la utilidad Y (en \$) que produce cada llanta está dada por la relación: $Y = 0.2X + 100$, ¿cuál es la probabilidad de que la utilidad sea mayor que 8,900\$?

b) Determinar el número de tales llantas que debe adquirir una empresa de transporte para conseguir una utilidad promedio de al menos \$7541 con probabilidad 0.996. Rp. a) 0.0228, b) $n = 100$

12. Un proceso automático llena bolsas de café cuyo peso neto tiene una media de 250 gramos y una desviación estándar de 3 gramos. Para controlar el proceso, cada hora se pesan 36 bolsas escogidas al azar; si el peso neto medio está entre 249 y 251 gramos se continúa con el proceso aceptando que el peso neto medio es 250 gramos y en caso contrario, se detiene el proceso para reajustar la máquina.

a) ¿Cuál es la probabilidad de detener el proceso cuando el peso neto medio realmente es 250?.

b) ¿Cuál es la probabilidad de aceptar que el peso neto promedio es 250 cuando realmente es de 248 gramos?.

Rp. a) 0.0456, b) 0.0228

13. En cierta población de matrimonios el peso en kilogramos de esposos y esposas se distribuye normalmente $N(80,100)$ y $N(64,69)$ respectivamente y son independientes. Si se eligen 25 matrimonios al azar de esa población calcular la probabilidad de que la media de los pesos sea a lo más 137 Kg.

Rp. 0.0036.

14. Una empresa vende bloques de mármol cuyo peso se distribuye normalmente con una media de 200 kilogramos.

a) Calcular la varianza del peso de los bloques, si la probabilidad de que el peso esté entre 165 Kg. y 235 Kg. es 0.9876.

b) ¿Que tan grande debe ser la muestra para que haya una probabilidad de 0.9938 de que el peso medio de la muestra sea inferior a 205 Kg.?

Rp. a) σ=14, b) n=49

15. La duración en horas de una marca de tarjeta electrónica se distribuye exponencialmente con un promedio de 1000 horas.

a) Hallar el tamaño n de la muestra de manera que sea 0.9544 la probabilidad de que su media muestral esté entre 800 y 1200 horas.

b) Si se obtiene una muestra aleatoria de 100 de esas tarjetas calcular la probabilidad que la duración media de la muestra sea superior a 1,100 horas. Rp. a) n = 100, b) 0.1587.

16. Un proceso para llenar cerveza en botellas de 620 ml. sufre una pérdida en el contenido que tiene una media de 5 ml. y una desviación estándar de 1.2 ml.. Se escogen al azar 36 de tales botellas. Si la meda de la muestra está entre 4.5 y 5.5 ml. se acepta que μ=5 ml., en caso contrario; se rechaza que μ=5. ¿Cuál es la probabilidad de aceptar que μ=5 cuando realmente es μ=4.8 ml.?.

Rp. 0.9330.

17. Una empresa comercializa fardos de algodón cuyo peso X se distribuye normalmente con una media de 250 Kg. y una desviación estándar de 4 Kg. El costo por fardo es dado por $Y = aX + 52$. Hallar el valor de a si se quiere que la media de los costos de 4 fardos sea mayor que $3,100 con probabilidad 0.0228.

Rp. 12

18. Definimos la variable aleatoria "error muestral", por: $|\overline{X} - \mu|$. De todas las muestras de tamaño 36 escogidas al azar de la población $N(\mu,324)$,

a) ¿qué porcentaje tendrán un error muestral mayor de 4.5?.
b) ¿para qué valor de k el 95% tienen error muestral no mayor que k?.

Rp. a) 0.1336, b) $k = 5.88$.

19. El costo de producción en dólares de un objeto es 100 veces el valor numérico de su longitud. Suponga que la longitud en metros del objeto es una variable aleatoria con distribución normal $N(0.012, 1.44\times10^{-4})$.

a) ¿Cuál es la distribución del costo medio por objeto si se toman al azar n ($n \geq 2$) objetos?.
b) Si el precio de venta de cada objeto es $2.00, calcular la probabilidad de que la utilidad promedio por objeto de 36 objetos tomados al azar, sea a lo más $0.5.

Rp. a) $N(1.2, 1.44/n)$, b) $\overline{U}_{36} \cong N(0.8, 0.04)$, $P[\overline{U}_{36} \leq 0.5]=0.0668$.

20. Un analista de investigación de mercado toma una muestra aleatoria de 36 clientes de una tienda, de un conjunto de 400 clientes que adquirieron un cupón especial. El monto de las compras mensuales de los 400 clientes constituye una población finita con una media de 2,500 dólares y una desviación estándar de $660. ¿Cuál es la probabilidad de que la media de la muestra supere los $2765?.

Rp. población finita, 0.0059

21. Un auditor quiere tomar una muestra aleatoria de una población que consiste de 10,000 cuentas por cobrar, donde $\sigma = \$2000$. ¿De que tamaño debe escoger la muestra si se quiere tener una probabilidad del 95% de que la diferencia entre la media muestral y la media poblacional no exceda el valor $192?

Rp. Población finita, $n \cong 400$

22. La calificación en una prueba de aptitud es una variable aleatoria X que tiene distribución normal con media igual a 100.

a) Si se supone que la desviación estándar de todas las calificaciones es $\sigma = 15$, ¿cuántas calificaciones se deben escoger para que la media muestral esté en el intervalo de 90.2 a 109.8 con probabilidad 0.95?.

b) Si se escogen al azar 16 calificaciones y se encuentra que la desviación estándar $\hat{s} = 12$, ¿cuál es la probabilidad de que la media muestral se encuentre entre 92.194 y 104.023?. Rp. a) $n = 9$, b) 0.89.

23. El gerente de producción afirma que las baterías que produce duran en promedio tres años. En el control de calidad se verifican 16 baterías y si el valor de t calculado: $t_c = (\bar{x} - 3)/(\hat{s}/\sqrt{n})$ está entre $-t_{0.05}$ y $t_{0.05}$, el fabricante está satisfecho con su afirmación. ¿Qué conclusión sacará el fabricante si la muestra da una media de 3.8 años y una desviación estándar $\hat{s} = 1.5$ años?. Suponga que la duración de las baterías tiene distribución normal.

Rp. $t_c=2.13 \notin [-1.753, 1.753]$, es un producto mejor de lo afirmado.

Una Proporción.

24. Se estima que el 40% de los votos de los electores de la ciudad favorecen al candidato Sr. Díaz.

a) Si se selecciona una muestra aleatoria de 600 electores de la ciudad, ¿qué probabilidad hay de que la proporción muestral de votos a favor del Sr. Díaz esté entre 37% y 45%?

b) ¿Qué tamaño de muestra se debería escoger si se quiere tener una probabilidad igual a 0.97 de que la proporción de votos a favor del Sr. Díaz en la muestra no se diferencie de su proporción estimada en más del 2%?.

Rp. a) $P[-1.5 \leq Z \leq 2.5]=0.9270$, b) $n=2825.34=2826$.

25. Una empresa que hace estudios de mercado quiere obtener una muestra aleatoria suficientemente grande de manera que la probabilidad de que la proporción obtenida a favor de un cierto producto resulte inferior al 35% sea igual a 0.0062.

a) Calcular el tamaño de la muestra a tomar si se supone que la verdadera proporción a favor del producto es $p = 0.4$.

b) Con el tamaño de muestra calculado en a) y si se supone verdadero el valor del parámetro $p = 0.2$, determinar el intervalo $[a,b]$ centrado en p tal que $\bar{P} \in [a,b]$ con probabilidad 0.95 Rp. a) $n = 600$, b) [0.1608, 0.2392]

26. Un fabricante afirma que a lo más el 2% de todas las piezas producidas son defectuosas. Al parecer esta información es exagerada, por lo que se selecciona una muestra aleatoria de 400 de tales piezas. Si la proporción muestral de

defectuosos es mayor que 3% se rechaza la afirmación, en caso contrario se acepta la afirmación.

a) ¿Cuál es la probabilidad de rechazar la afirmación cuando realmente el 2% de todas las piezas producidas son defectuosas?

b) ¿Cuál es la probabilidad de aceptar la afirmación cuando realmente el 4% de todas las piezas producidas son defectuosas?

Rp. a) 0.0764. b) 0.1539

27. El director de la bolsa de trabajo de la universidad afirma que el 60% de los egresados consigue empleo con una remuneración mayor a los $500. Para comprobar esta afirmación se escoge una muestra aleatoria de 600 egresados de la universidad. Si 330 o más pero no más de 390 de la muestra consiguen trabajo con remuneración mayor a los $500, se aceptará la afirmación. En caso contrario se rechazará tal afirmación.

a) ¿Cuál es la probabilidad de rechazar la afirmación cuando ésta es realmente verdadera.?.

b) ¿Cuál es la probabilidad de aceptar la afirmación cuando realmente el 70% de todos los egresados consiguen trabajo con remuneración mayor a los $500?. Rp. a) 0.0124. b) 0.0038.

28. Para controlar la calidad en un proceso de producción de cierto bien de consumo, se seleccionan al azar 46 unidades del bien cada día. Si la proporción de objetos defectuosos en la muestra es al menos $\overline{p}_0$, se detiene el proceso, de otro modo se continua con el proceso. Determine aproximadamente el valor de $\overline{p}_0$ para que con probabilidad de 0.9332 no se continúe con el proceso, cuando la producción total contenga 8% de objetos defectuosos.

Rp. 0.02.

29. Un nuevo producto va a salir al mercado si por lo menos el $\overline{p}_0$ (100%) de n personas encuestadas, aceptan el producto. Calcular los valores de n y $\overline{p}_0$ de manera que haya una probabilidad de 0.1112 de que el producto no saldrá al mercado cuando realmente el 58% lo aceptan y una probabilidad de 0.0228 de que el producto saldrá al mercado cuando realmente el 50% lo aceptan.

Rp. $n \cong 400$, $\overline{p}_0 \cong 0.55$.

30. Por experiencia el departamento de créditos de una tienda comercial sabe que sus ventas se pagan con: dinero en efectivo, con cheque o al crédito, con probabilidades respectivas; 0.3, 0.3, y 0.4.. La probabilidad de que una venta sea por más de $50 es igual a 0.2 si ésta es en efectivo, es igual a 0.9 si ésta es con cheque y es igual a 0.6 si ésta es al crédito.

Si se escoge una muestra aleatoria de 256 personas que ingresan a la tienda, ¿cuál es la probabilidad de que el porcentaje de personas que hayan comprado por más de $50 sea al menos 50%?.

Rp. 0.9881.

31. De 3000 empleados de una empresa se escoge una muestra aleatoria de 300 empleados para una encuesta sobre condiciones laborales. ¿Cuál es la probabilidad de que la proporción muestral a favor de las condiciones laborales esté comprendido en el intervalo 0.76 y 0.84, si se estima en 80% del total de empleados el porcentaje a favor de las condiciones laborales?.

Rp. 0.9328

32. Una empresa encuestadora debe seleccionar una muestra aleatoria de una población que consiste de 3000 electores para una encuesta de opinión. La empresa estima en 30% del total, el porcentaje a favor de cierto candidato. ¿De que tamaño debe escoger la muestra si se quiere tener una probabilidad del 95% de que la diferencia de la proporción a favor del candidato en la muestra y en la población no exceda el valo 0.0492?

Rp. $n \cong 300$

Varianzas

33. Si $X_1, X_2, ..., X_8$ son ocho variables aleatorias independientes y distribuidas cada una normal $N(10,32)$, calcular la probabilidad de que la varianza muestral $S^2 = \Sigma(X_i - \overline{X})^2/8$ sea menor o igual que 56.28. Rp. 0.95.

34. Calcular la probabilidad de que una muestra aleatoria de tamaño 13 escogida de una población normal con varianza $\sigma^2 = 4$ tenga una varianza muestral $\hat{S}^2$,
a) menor que 7.01, b) entre 1.19 y 2.1.

Rp. a) 0.95, b) 0.09.

35. Si $X_1, X_2, ..., X_9$ son 9 variables aleatorias independientes y con distribución normal, $N(8,4)$, calcular la probabilidad $P[1.09 \leq \hat{S}^2 \leq 10.045,\ 7 \leq \overline{X} \leq 9]$ ($\overline{X}$ y $\hat{S}^2$ son independientes). Rp. 0.8361.

36. Utilizando la tabla de la distribución F hallar:
a) $F_{0.95,10,15}$, b) $F_{0.99,15,9}$, c) $F_{0.05,30,8}$, d) $F_{0.01,15,9}$

Rp. a) 2.54, b) 4.96, c) 0.4405, d) 0.257.

37. Dos muestras aleatorias independientes de tamaños 21 y 9 respectivamente se toman de una misma población que está normalmente distribuida, ¿cuál es la probabilidad de que la varianza de la primera muestra sea al menos el cuádruple de la varianza de la segunda?.

Rp. 0.025

38. Sean. $X_1 \sim \chi^2(9)$, $X_2 \sim \chi^2(20)$ y $X = (X_1/9)/(X_2/20)$ hallar los valores a y b tales que:

$$P[a \leq X \leq b] = 0.925 \text{ y } P[X \leq a] = 0.05.$$

Rp. b=2.84, a=1/2.94=0.34

39. Sea $X_1, \ldots, X_{10}$ una muestra aleatoria escogida de una población normal $N(0,1)$,

a) Hallar la distribución de $F = (\sum_{i=1}^{10} X_i^2)/10 \Big/ (\sum_{i=1}^{5} X_i^2)/5$,

b) Calcular la probabilidad $P[F < 1/3.33]$

Rp. a) $F \sim F(10,5)$, b) 0.05

40. Sea X_1 y X_2 una muestra aleatoria escogida de una población normal $N(0,1)$.

a) Hallar la distribución de $F = \left[\dfrac{X_1 + X_2}{X_1 - X_2}\right]^2$,

b) Calcular la probabilidad $P[F < 161]$

Rp. a) $F \sim F(1,1)$, b) 0.95

Diferencia de dos medias

41. Para comparar la duración promedio (en meses) μ_1 y μ_2 de dos marcas de baterías B1 y B2 se escogen dos muestras aleatorias independientes de tamaños respectivos $n_1 = 32$ y $n_2 = 36$. Si la media muestral de B1 es mayor que la media muestral de B2 en mas de 2 meses, se acepta que $\mu_1 > \mu_2$.En caso contrario se acepta que $\mu_1 = \mu_2$. Calcular la probabilidad de aceptar que $\mu_1 > \mu_2$ cuando realmente $\mu_1 = \mu_2$. Suponga que las varianzas de las duraciones de B1 y B2 son respectivamente $\sigma_1^2 = 16$ y $\sigma_2^2 = 9$.

Rp. 0.0104

42. Una firma comercializadora afirma que el peso promedio (en gramos) μ_1 y μ_2 de dos marcas de café instantáneo C1 y C2, es el mismo. Para verificar la afirmación se escogen dos muestras aleatorias independientes de tamaños 36 sobres de cada marca.. Si la media muestral de C1 es mayor que la media muestral de C2 en mas de 0.5 gramos, se rechaza que $\mu_1 = \mu_2$. En caso contrario, se acepta que $\mu_1 = \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 = \mu_2$ cuando realmente $\mu_1 = \mu_2 + 2$?. Suponga que las varianzas de las poblaciones C1 y C2 son respectivamente $\sigma_1^2 = 9$ y $\sigma_2^2 = 4$

Rp. 0.0062

43. El jefe de compras está por decidir si comprar una marca A o una marca B de focos para la compañía. Para ayudarle a optar por una de ellas se escogen dos muestras aleatorias de tamaños $n_1 = 10$ y $n_2 = 9$ focos respectivamente de las marcas A y B, resultando, las desviaciones estándares respectivas $\hat{s}_1 = 200$ y $\hat{s}_2 = 150$. Si la diferencia entre las medias muestrales es mayor que 173 horas. se acepta que $\mu_1 \neq \mu_2$. En caso contrario, se acepta que $\mu_1 = \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 \neq \mu_2$ cuando realmente $\mu_1 = \mu_2$?. Se asume que la vida útil de ambas marcas tiene distribución normal con varianzas iguales.

Rp. $T \sim t(17)$, 0.05.

44. Para comparar los salarios que se pagan a los empleados en dos grandes empresas E1 y E2 se escogen dos muestras aleatorias independientes de tamaños $n_1 = 16$ y $n_2 = 13$ respectivamente de E1 y E2 resultando las desviaciones estándares respectivas $\hat{s}_1 = \$120$ y $\hat{s}_2 = \$55$. Si la diferencia entre las medias muestrales no es mayor que 65$, se acepta que $\mu_1 = \mu_2$. En caso contrario, se acepta que $\mu_1 \neq \mu_2$. ¿Cuál es la probabilidad de aceptar que $\mu_1 \neq \mu_2$ cuando realmente $\mu_1 = \mu_2$?. Se asume que los salarios en ambas empresas tienen una distribución normal con varianzas diferentes.

Rp. $T \sim t(22)$, 0.10

Diferencia de dos proporciones.

45. Dos programas de televisión A y B tienen como ratings (porcentaje de hogares donde se ve el programa) de 40 y 20 respectivamente. Se toma una muestra aleatoria de 300 hogares con T.V. durante la transmisión del programa A y otra de 100 hogares durante la transmisión de B, ¿cuál es la probabilidad de que los resultados muestren que el programa A tiene un rating mayor a la de B en 10%?.

Rp. 0.0207

46. Un fabricante afirma que el 30% de mujeres y el 20% de hombres prefieren su nuevo producto de aseo personal. Si se hace una encuesta a 200 hombres y 200 mujeres elegidos aleatoriamente, ¿con qué probabilidad la proporción muestral de mujeres menos la proporción muestral de hombres está en el intervalo [−19%, 19%]?

Rp. 0.9634

47. Se escoge una muestra de 600 electores que acaban de votar, entre las 9 a.m. y las 3 p.m. para estimar la proporción de votantes a favor de los candidatos A y B. En una encuesta hecha en la víspera se estimó en 30% y 35% los porcentajes a favor de A y B respectivamente. ¿Cuál es la probabilidad de que la proporción muestral de B exceda a la proporción muestral de A en al menos 10%?.

Rp. 0.0322.

Capítulo 9

ESTIMACION DE PARAMETROS.

9.1 Introducción

Al realizar una investigación estadística a menudo se sabe o se supone que la población (discreta o continua), de la cual se selecciona una muestra aleatoria, tiene una forma funcional específica $f(x)$ cuyo(s) parámetro(s) se intenta determinar. Si el parámetro a determinar es denotado por θ, entonces, la distribución de la población será denotada por $f(x, \theta)$.

Los métodos de inferencia estadística consisten en seleccionar una muestra aleatoria de la población, de manera que a partir de la información que se obtenga de la muestra:

1) Determinar el valor del parámetro desconocido θ, ó

2) Decidir si θ, ó alguna función de θ, es igual a algún valor preconcebido θ_0 de θ.

El primero de estos dos procedimientos se denomina ***estimación del parámetro*** θ. El segundo procedimiento se conoce como ***prueba de hipótesis del parámetro*** θ.

El método de estimación de un parámetro puede ser ***puntual*** o ***por intervalo***. En el primer caso, la estimación del parámetro θ es un número. Mientras que en el segundo caso la estimación incluye un intervalo en el que están comprendidos los valores del parámetro.

9.2 Estimación puntual de parámetros

Definición. Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño *n* seleccionada de una población cuya distribución es $f(x, \theta)$, siendo θ el parámetro. Se denomina ***estimador puntual*** del parámetro θ a cualquier estadística $\hat{\Theta} = H(X_1, X_2,..., X_n)$, cuyo valor $\hat{\theta} = H(x_1, x_2,..., x_n)$ proporcionará una ***estimación*** del parámetro en cuestión.

Un estimador puntual del parámetro θ es pues, una *variable aleatoria* (función de la muestra) $\hat{\Theta}$, mientras que una estimación puntual es el *valor numérico* $\hat{\theta}$ del estimador.

Por dar un ejemplo, un **estimador** puntual de la media poblacional θ, es la estadística media muestral (variable aleatoria) $\hat{\Theta} = \overline{X}$, cuyo valor numérico $\hat{\theta} = \overline{x}$ es la **estimación** puntual del parámetro θ.

No toda función de la muestra es un *buen estimador* del parámetro, un buen estimador, es aquel que está más cerca del parámetro que se estima. Para que un estimador puntual sea bueno debe tener ciertas propiedades. Una de estas propiedades es que sea *insesgado*, propiedad conocida también como *no-sesgado*, *imparcial*, o ***sin vicio***.

9.2.1. Estimador insesgado

Definición. Se dice que la estadística $\hat{\Theta} = H(X_1, X_2,..., X_n)$ es un **estimador insesgado** del parámetro θ si $E(\hat{\Theta}) = \theta$. En caso contrario, se dice que es estimador sesgado.

Si la estadística $\hat{\Theta} = H(X_1, X_2,..., X_n)$ es un estimador insesgado del parámetro θ, entonces, su valor $\theta = H(x_1, x_2,..., x_n)$ es la **estimación insesgada** del parámetro θ.

EJEMPLO 9.1.

Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño *n* extraída de una población cualquiera $f(x, \mu, \sigma^2)$, (discreta o continua). Entonces,

a) La media muestral $\overline{X}$ es un estimador insesgado de la media poblacional μ, ya que

$$E(\overline{X}) = \mu.$$

El valor $\bar{x}$ de $\overline{X}$ es la estimación insesgada de μ.

b) La proporción muestral $\overline{P}$ es un estimador insesgado de la proporción de éxitos p de una población binomial, por que,

$$E(\overline{P}) = p.$$

c) La varianza muestral

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

es un **estimador sesgado** de la varianza poblacional σ^2, ya que

$$E(S^2) = \frac{n-1}{n}\sigma^2.$$

Sin embargo, la estadística,

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

es un estimador insesgado de la varianza poblacional σ^2, por que,

$$E(\hat{S}^2) = \sigma^2.$$

NOTA. El cálculo la varianza por $\hat{S}^2$ es denominado **método de muestra** mientras que el cálculo de la varianza por S^2 es denominado **método de población**. Observar que cuando n es suficientemente grande el coeficiente $(n-1)/n$ tiende a uno.

EJEMPLO 9.2.

a) La diferencia $\overline{X}_1 - \overline{X}_2$ de dos medias muestrales es estimador insesgado de la diferencia $\mu_1 - \mu_2$ de dos medias poblaciones. Ya que,

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2.$$

b) La diferencia de dos proporciones muestrales $\overline{P}_1 - \overline{P}_2$ es un estimador insesgado de la diferencia de dos proporciones de éxitos binomiales $p_1 - p_2$. Puesto que,

$$E(\overline{P}_1 - \overline{P}_2) = p_1 - p_2.$$

9.2.2. Estimador eficiente

Definición. Si hay dos o más estimadores puntuales insesgados de un parámetro θ, se denomina estimador **más eficiente** a aquel estimador que tenga *menor varianza*.

EJEMPLO 9.3.

Sea X_1, X_2, X_3, X_4 una muestra aleatoria de cualquier población con distribución $f(x, \mu, \sigma^2)$. Dados los estimadores del parámetro μ:

$$\text{a) } \hat{\Theta}_1 = \frac{X_1 + X_2 + X_3 + X_4}{4}, \quad \text{y} \quad \text{b) } \hat{\Theta}_2 = \frac{4X_1 - X_3 + X_4}{4}$$

verificar que la media de la muestra; es el estimador más eficiente de μ.

SOLUCION.

Ambos estimadores son insesgados. En efecto,

$$E(\hat{\Theta}_1) = \frac{4\mu}{4} = \mu \quad \text{y} \quad E(\hat{\Theta}_2) = \frac{4\mu - \mu + \mu}{4} = \frac{4\mu}{4} = \mu$$

Por otra parte, las varianzas respectivas son:

$$V(\hat{\Theta}_1) = \frac{4\sigma^2}{16} = \frac{\sigma^2}{4} \quad \text{y} \quad E(\hat{\Theta}_2) = \frac{16\sigma^2 + \sigma^2 + \sigma^2}{16} = \frac{18\sigma^2}{16}$$

Luego, la estadística $\hat{\Theta}_1$ que está definida como la media de la muestra; es el estimador más eficiente de μ.

Uno de los métodos para determinar estimadores puntuales es el de *máxima verosimilitud* que se describe a continuación.

9.2.3. Método de máxima verosimilitud

Supongamos que una población X está distribuida como $f(x, \theta)$, en donde θ es el parámetro que tratamos de estimar.

El procedimiento para determinar el estimador de máxima verosimilitud es como sigue:

1) Elegir una muestra aleatoria $X_1, X_2, \ldots, X_n$ de la población y determinar la distribución conjunta de la muestra en sus valores observados respectivos $x_1, x_2, \ldots, x_n$. Esta función del parámetro θ conocida también como **función de verosimilitud** está dada por:

$$L(\theta) = f(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) f(x_2, \theta) ... f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

2) El valor de θ que maximiza a la función $L(\theta)$, es la **estimación** de máxima verosimilitud (EMV) de θ. Este valor denotaremos por

$$\hat{\theta} = H(x_1, x_2, ..., x_n).$$

La estadística correspondiente $\hat{\Theta} = H(X_1, X_2, ..., X_n)$ es el **estimador** de máxima verosimilitud de θ.

3) A menudo se usa $L = \ln(L(\theta))$. En este caso el valor de θ que maximiza a $L(\theta)$ es la solución $\hat{\theta}$ de la ecuación:

$$\frac{dL}{d\theta} = 0.$$

4) Si la distribución de probabilidad de la población contiene k parámetros $\theta_1, \theta_2, ..., \theta_k$, la función de verosimilitud está dada por:

$$L(\theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2, ..., \theta_k).$$

La estimación de máxima verosimilitud de cada parámetro θ_i es la solución $\hat{\theta}_i$ de la ecuación respectiva:

$$\frac{\partial L}{\partial \theta_1} = 0, \ \frac{\partial L}{\partial \theta_2} = 0, ..., \ \frac{\partial L}{\partial \theta_k} = 0$$

donde $L = \ln(L(\theta_1, \theta_2, ..., \theta_k))$.

EJEMPLO 9.4.

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n, seleccionada de una población con distribución binomial $B(1,p)$. Hallar el estimador del parámetro p usando el método de máxima verosimilitud.

SOLUCION.

La distribución de probabilidad de cada variable aleatoria X_i es:

$$f(x_i, p) = p^{x_i}(1-p)^{1-x_i}, \ x_i = 0, 1.$$

La función de verosimilitud de la muestra aleatoria es entonces,

$$L(p)=\prod_{i=1}^{n} f(x_i,p)=(p)^{\Sigma x_i}(1-p)^{n-\Sigma x_i} \qquad (*)$$

Luego, $$L=\ln(L(p))=(\ln(p))\sum_{i=1}^{n} x_i+(\ln(1-p))\left(n-\sum_{i=1}^{n} x_i\right).$$

Derivando la función L con respecto a p e igualando a cero, da:

$$\frac{dL}{dp}=\frac{\sum_{i=1}^{n} x_i}{p}-\frac{n-\sum_{i=1}^{n} x_i}{1-p}=0.$$

De donde resulta que:

$$\hat{p}=\frac{\sum_{i=1}^{n} x_i}{n}=\overline{p}.$$

Luego, la proporción muestral $\overline{p}$ es la estimación ($\overline{P}$ es el estimador) de máxima verosimilitud de la proporción poblacional p, siempre que:

$$\sum_{i=1}^{n} x_i \neq 0, \quad \text{y} \quad \sum_{i=1}^{n} x_i \neq n.$$

Observar que:

si $\sum_{i=1}^{n} x_i = 0$, entonces, de (*) resulta que el E.M.V. de p es $\hat{p}=0$.

si $\sum_{i=1}^{n} x_i = n$, entonces, de (*) resulta que el E.M.V. de p es $\hat{p}=1$.

EJEMPLO 9.5.

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria escogida de una población con distribución $N(\mu,\sigma^2)$. Hallar el estimador de la media μ y la varianza σ^2 de la distribución, usando el método de máxima verosimilitud.

SOLUCION.

La distribución de probabilidades de la población normal de parámetros μ y σ^2 asociada a cada variable aleatoria X_i está dada por:

$$f(x_i,\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

La función de verosimilitud es:

$$L(\mu,\sigma^2)=\left[f(x_i,\mu,\sigma^2)\right]^n=\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right]^n.$$

$$L(\mu,\sigma^2)=(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

Luego,

$$L=\ln(L(\mu,\sigma^2))=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Derivando la función L con respecto a μ e igualando a cero da:

$$\frac{dL}{d\mu}=-0+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

de donde resulta: $$\hat{\mu}=\frac{\sum_{i=1}^{n}x_i}{n}=\bar{x}.$$

Luego, la media muestral $\bar{x}$ es la estimación ($\bar{X}$ es el estimador) de máxima verosimilitud de la media poblacional μ.

Por otro lado, derivando la función L con respecto a σ^2 e igualando a cero da:

$$\frac{dL}{d\sigma^2}=-\frac{1}{2}\frac{n}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

de donde resulta: $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = s^2.$$

Luego, la varianza muestral s^2 es la estimación (S^2 es el estimador) de máxima verosimilitud de σ^2.

NOTA. Los estimadores de máxima verosimilitud tienen la ***propiedad de invarianza***, en el sentido de que **si $\hat{\theta}$ es E.M.V del parámetro θ, entonces, cualquier expresión $g(\hat{\theta})$ es E.M.V. de $g(\theta)$**. En el ejemplo 9.5, la varianza muestral s^2 es la estimación de máxima verosimilitud de la varianza poblacional σ^2, luego, la desviación estándar muestral s es la estimación de máxima verosimilitud de la desviación estándar σ.

EJEMPLO 9.6.

Para estimar la vida media de un tipo de componente electrónica se selecciona una muestra aleatoria de 10 unidades, se les somete a prueba y se encuentra que 6 de ellas siguen funcionando después de 3,000 horas. Suponiendo que la vida útil de las componentes es una variable aleatoria T con distribución exponencial de parámetro β, estimar la vida media de tales componentes producidas.

SOLUCION.

Sea p la proporción poblacional de todas las componentes que siguen funcionando después de 3,000 horas. Entonces,

$$p = P[T > 3000] = \int_{3,000}^{+\infty} \beta e^{-\beta t} dt = e^{-3000\beta}$$

De donde resulta: $$\hat{p} = e^{-3000\hat{\beta}}.$$

Por otra parte, la estimación por máxima verosimilitud de la proporción p es (ejemplo 9.4):

$$\hat{p} = \bar{p} = 6/10.$$

Por tanto, utilizando la propiedad de invarianza en $\hat{p} = e^{-3000\hat{\beta}}$. resulta:

$$\hat{\beta} = -\frac{1}{3,000}\ln(\hat{p}) = -\frac{1}{3,000}\ln(\frac{6}{10}) \quad \text{y} \quad \hat{\bar{x}} = \frac{1}{\hat{\beta}} = 5.872.85 \text{ horas}$$

EJERCICIOS.

1. De una población $f(x,\mu,\sigma^2)$ se escogen dos muestras aleatorias independientes de tamaños n_1 y n_2. Sean $\overline{X}_1$ y $\overline{X}_2$, y $\hat{S}_1^2$ y $\hat{S}_2^2$ sus medias y varianzas respectivas.

a) Si $\overline{X}=(n_1\overline{X}_1+n_2\overline{X}_2)/(n_1+n_2)$, ¿es la estadística $\overline{X}$ un estimador insesgado del parámetro μ?.

b) Si $\hat{S}_c^2=\dfrac{(n_1-1)\hat{S}_1^2+(n_2-1)\hat{S}_2^2}{n_1+n_2-2}$, ¿es la estadística $\hat{S}_c^2$ un estimador insesgado del parámetro σ^2?.

Rp. a) si, b) si.

2. Si $\overline{P}_1$ y $\overline{P}_2$ son las proporciones de dos muestras de tamaños n_1 y n_2 escogidas de una población Bernoullí $B(1, p)$, verifique que la estadística $\hat{P}=\dfrac{n_1\overline{P}_1+n_2\overline{P}_2}{n_1+n_2}$ es un estimador insesgado del parámetro p.

3. Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras independientes de tamaños n_1 y n_2 respectivamente escogidas de una población X de Poisson con parámetro λ.

a) Probar que la estadística $\hat{\theta}=\dfrac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$ es un estimador insesgado de del parámetro λ.

b) Hallar la varianza del estimador.

Rp. a) $E(\hat{\theta})=\lambda$, b) $V(\hat{\theta})=\dfrac{\lambda}{n_1+n_2}$ ..

4. La duración, en horas, de cierta clase de foco sigue una distribución exponencial con media desconocida θ horas. Se toma una muestra de un sólo foco al azar y se mide su duración X en horas. Si con X se estima θ, ¿se podría decir que X es un estimador insesgado de θ?.

Rp. Si por que $X\sim \text{Exp}(1/\theta)$, entonces, $E(X)=(1/\theta)^{-1}=\theta$

5. Dos métodos diferentes e independientes dieron lugar a dos estimadores insesgados $\hat{\theta}_1$ y $\hat{\theta}_2$ del parámetro θ. Las desviaciones estándares de estos estimadores son 0.4 y 0.6 respectivamente. Los estimadores son combinados de la siguiente manera:

$$\hat{\theta} = r\hat{\theta}_1 + (1-r)\hat{\theta}_2 \quad 0 < r < 1.$$

Hallar el valor de r que haga mínima la varianza del estimador $\hat{\theta}$.

Rp. r=0.6923.

6. Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n de una población de Bernoullí $B(1, p)$. De las siguientes estadísticas:

$$\hat{\Theta}_1 = \frac{\sum_{i=1}^{n} X_i - X_k}{n-1}, \qquad \hat{\Theta}_2 = \frac{\sum_{i=1}^{n} X_i^2}{n}$$

a) ¿Cuáles son estimadores insesgados del parámetro p?.

b) ¿Cuál de ellas es de varianza mínima?.

Rp. a) ambos, b) 2do.

7. Sea $X_1, X_2, ..., X_{50}$ una muestra aleatoria de tamaño 50 escogida de una población con distribución geométrica de parámetro p, $0 < p < 1$,

$$P[X = x] = p(1-p)^x, \quad x = 0,1,2,...$$

a) Determinar el estimador de máxima verosimilitud para p.

b) Estimar p, si $x_1 + x_2 + \cdots + x_{50} = 100$.

Rp. a) $1/(\bar{x}+1)$, b) 1/3.

8. Sea $X_1, X_2, ..., X_{20}$ una muestra aleatoria de tamaño 20 escogida de una población con distribución binomial de parámetro p, $0 < p < 1$,

$$P[X = x] = C_x^2 p^x (1-p)^{2-x}, \quad x = 0, 1, 2$$

Determinar el estimador de máxima verosimilitud para p, si en la muestra el valor 0 ocurre 4 veces, el valor 1 ocurre 9 veces y el valor 2 ocurre 7 veces.

Rp. $L(p) = 2^9 p^{23} (1-p)^{17}$, $\hat{p} = 23/40$.

9. El número de ventas diarias de cierta mercadería es una variable aleatoria X de Poisson con un promedio de λ ventas por día.

a) Si $X_1, X_2, ..., X_{50}$, son las ventas de 50 días, estimar λ, por el método de máxima verosimilitud.

b) Si en los 50 días se han hecho 30 ventas de tal mercadería, estimar el promedio λ de ventas diarias.

Rp. a) $\hat{\lambda} = \bar{x}$, b) 30/50

10. De una población de variable aleatoria continua X se extrae una muestra aleatoria $X_1, X_2, ..., X_n$, y se define la variable aleatoria Bernoullí:

$$Y = \begin{cases} 1 & \text{si } X_i > 0 \\ 0 & \text{si } X_i \le 0 \end{cases}$$

a) Usando el método de máxima verosimilitud estimar la proporción p de todos los valores positivos, esto es, estimar, $p = P[X > 0]$.

b) Estime el valor de p si una muestra aleatoria de tamaño 80 de X ha dado 64 valores positivos y 16 valores negativos.

c) Si $X \sim N(\mu, 0.04)$, utilizando a) y b), calcular aproximadamente el valor de μ.

Rp. a) $\hat{p} = \sum y_i / n$, b) 0.8, c) $\mu = 0.168$.

11. El tiempo, en meses, que dura una componente electrónica es una variable aleatoria T de distribución exponencial con parámetro β. Para estimar β se prueban 30 componentes y se encuentra que 18 fallan antes de los 6 meses.

a) Utilizando el método de máxima verosimilitud, estimar la proporción de todas las componentes que fallan antes de los 6 meses.

b) Utilice el resultado de a) para estimar por máxima verosimilitud de β.

Rp. a) $\hat{p} = 18/30$, b) $\hat{p} = 1 - e^{-6\hat{\beta}}$, $\hat{\beta} = -\ln(12/30)/6$

12. La longitud de cierto tipo de objeto producidos por una máquina, puede estar por arriba o por abajo de la medida estándar de 2 pulgadas. Suponga que tal longitud tiene distribución normal $N(\mu, 0.0025)$.

a) Utilizando el método de máxima verosimilitud estime la proporción p de todos los objetos cuya longitud está por arriba de 2 pulgadas.

b) Si en una muestra de 1,000 de tales objetos se encontró que 992 tenían longitud por arriba de 2 pulgadas, utilizando a) estime la media de la longitud de todos los objetos producidos.

Rp. a) $\hat{p} = \overline{p}$, b) $\overline{p} = 0.992 = P[Z > (2 - \hat{\mu})/0.05]$, $\hat{\mu} = 2.1205$

13. Una máquina produce objetos cuyo peso en gramos tiene distribución normal $N(30, \sigma^2)$, con σ desconocido. Los objetos son defectuosos si el peso es menor que 26 o mayor que 34 gramos. Para estimar σ se pesa un objeto cada vez hasta que un defectuoso sea obtenido. Hallar el estimador de máxima verosimilitud de σ si en un control el primer defectuoso se halló en la décima prueba.

Rp. $1 - \hat{p} = P[-4/\hat{\sigma} \le Z \le 4/\hat{\sigma}]$, $\hat{p} = 0.1$, $\hat{\sigma} = 2.43$.

9.3 Estimación de parámetros por intervalos

9.3.1 Intervalo de confianza

Una estimación de punto no nos dice cuán próximo está la estimación al parámetro que se estima, por lo tanto, no es muy significativa, sin no se tiene alguna medida del error que se comete en la estimación. Es deseable pues tener cierto *grado de confianza* de que la estimación de punto se halle dentro de cierta variación.

La *estimación por intervalo* (propuesto por J. Neyman en 1937), es la estimación de un parámetro θ dentro de un intervalo de extremos cerrados $[a,b]$, donde los números a y b se obtienen a partir de la distribución de la estadística que estima puntualmente el parámetro; y a partir de los valores de la muestra.

Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño n escogida de una población $f(x, \theta)$, cuyos valores experimentales (o datos) respectivos son $x_1, x_2,..., x_n$. Sea además, la variable aleatoria $\hat{\Theta} = H(X_1, X_2,..., X_n)$ una estadística para estimar el parámetro θ cuya distribución de probabilidad sea conocida. Si dado el número $1-\alpha$, y si a partir de la distribución de $\hat{\Theta}$ se pueden encontrar las variables aleatorias $A = H_1(X_1, X_2,..., X_n)$, y $B = H_2(X_1, X_2,..., X_n)$ tales que:

$$P[A \le \theta \le B] = 1-\alpha$$

entonces, se dice que *el intervalo aleatorio* $[A, B]$ es el *intervalo estimador* del parámetro θ con el *grado o nivel de confianza* de $(1-\alpha)\times 100\%$, o que $\theta \in [A, B]$ con probabilidad $1-\alpha$.

Además, si $a = H_1(x_1, x_2,..., x_n)$, y $b = H_2(x_1, x_2,..., x_n)$, son los valores numéricos que resultan de sustituir los valores de la muestra en las *estadísticas* A y B respectivamente, entonces, se dice que el *intervalo numérico* $[a,b]$ es *el intervalo de confianza del* $(1-\alpha)\times 100\%$ para θ, o que $\theta \in [a,b]$ con *nivel o grado de confianza del* $(1-\alpha)\times 100\%$.

La diferencia, pues, entre los intervalos $[A, B]$ y $[a,b]$, es que el primero es un intervalo aleatorio y por lo tanto tiene validez afirmar que la probabilidad que contenga al parámetro θ es igual a $1-\alpha$. Mientras que el segundo es un intervalo numérico fijo, y en este caso, *no tiene validez afirmar* que la probabilidad $P[a \le \theta \le b] = 1-\alpha$.

La **interpretación** del intervalo de confianza es como sigue: Si a partir de los datos de una muestra aleatoria de tamaño n, hemos construido el intervalo $a \le \theta \le b$ con grado de confianza, por ejemplo, del 95% para el parámetro θ, entonces, si se seleccionan repetidamente 100 muestras de tamaño n, tendremos 100 intervalos semejantes al intervalo $[a,b]$, y se confía que 95 de estos 100 intervalos contengan el parámetro θ.

La probabilidad $1-\alpha$, o el porcentaje $(1-\alpha)\times 100\%$ es denominado *el grado (o nivel) de confianza*. Sus valores más utilizados son 0.95, 98, 0.99 entre otros. Al número α se le denomina también *riesgo de estimación* por intervalo.

A los números a y b se les denomina los *limites de confianza* o de *tolerancia* del parámetro θ. El número a es el *límite inferior de confianza* y el número b es el *límite superior de confianza*.

Por otra parte, si la estadística A_1 verifica:

$$P[A_1 \le \theta] = 1-\alpha$$

se concluye que el intervalo $[a_1,+\infty[$ es un *intervalo de estimación unilateral* del parámetro θ del $(1-\alpha)\times 100\%$, donde a_1 es el valor de A_1 que se obtiene a partir de la muestra.

Similarmente, si la estadística B_1 verifica:

$$P[\theta \le B_1] = 1-\alpha$$

se concluye que el intervalo $]-\infty, b_1]$ es un *intervalo de estimación unilateral* del parámetro θ del $(1-\alpha)\times 100\%$, donde b_1 es el valor de B_1 que se obtiene a partir de la muestra.

9.4 Intervalo de confianza para la media μ

9.4.1 Intervalo de confianza para la media μ: Varianza σ² supuesta conocida

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n seleccionada de una población normal (o de cualquier otro tipo, siendo n grande) con media μ y varianza σ^2 supuestamente conocida.

El mejor estimador puntual del parámetro μ es la media muestral $\overline{X}$.

Se puede utilizar, entonces, la distribución muestral de la media $\overline{X}$ para determinar el intervalo de confianza del parámetro μ.

Si la *población es normal* $N(\mu, \sigma^2)$, entonces, la distribución del estadístico $\overline{X}$ es normal $N(\mu, \sigma^2/n)$ para cualquier valor de n ($n \geq 2$).

Si la *población no es normal*, pero tiene media μ y varianza σ^2 finitas, entonces, siempre que el tamaño n de la muestra sea suficientemente grande, ($n \geq 30$), por el teorema del límite central, la distribución de $\overline{X}$ es *aproximadamente* normal $N(\mu, \sigma^2/n)$.

Por tanto, según sea el caso, la distribución de la variable aleatoria:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

es *exactamente* (o *aproximadamente*) normal $N(0,1)$.

Luego, dado el valor $1-\alpha$ (o en %), en la distribución de Z, se pueden determinar los valores $\mp z_{1-\alpha/2}$ (figura 9.1) tales que:

$$P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1-\alpha.$$

Sustituyendo $Z = (\overline{X} - \mu)\sigma/\sqrt{n}$, se tiene,

$$P\left[-z_{1-\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1-\alpha.$$

De donde resulta,

$$P\left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha.$$

$$P[A \leq \mu \leq B] = 1-\alpha.$$

donde $A = \overline{X} - z_{1-\alpha/2}\,\sigma/\sqrt{n}$ y $B = \overline{X} + z_{1-\alpha/2}\,\sigma/\sqrt{n}$ son variables aleatorias. Esto es, si $\overline{X}$ es estimador de μ, **se tiene la probabilidad** $1-\alpha$ **de que el *intervalo* (*aleatorio* o *estimador*)** $[A, B]$ contenga al parámetro μ.

Luego,

Si $\bar{x}$ es el valor de la media $\bar{X}$ para una muestra aleatoria de tamaño n escogida de una población con varianza σ^2 supuesta conocida, el intervalo de confianza del $(1-\alpha)\times 100\%$ para μ es:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

El valor $z_{1-\alpha/2}$ se busca en la tabla normal $N(0,1)$, tal que $P[Z \leq z_{1-\alpha/2}] = 1-\alpha/2$.

La ilustración, es la figura 9.1, en la que los valores $a = \bar{x} - z_{1-\alpha/2}\,\sigma/\sqrt{n}$ y $b = \bar{x} + z_{1-\alpha/2}\,\sigma/\sqrt{n}$ son los *límites de confianza* de μ, inferior y superior, respectivamente.

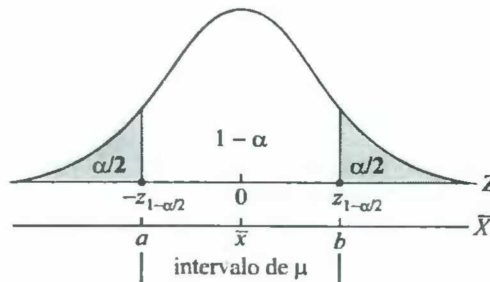


Figura 9.1: Intervalo de estimación para μ (con estadística Z.)

Interpretación. Si se seleccionan repetidamente 100 muestras de tamaño n, y calculamos las medias de cada una de ellas, tendremos 100 intervalos semejantes al intervalo $[a, b]$, y se confía que 95 de estos 100 intervalos contengan el parámetro μ y 5 de los 100 no lo contengan como se muestra en la figura 9.1b. Los puntos circulares en el centro de cada intervalo indican la estimación puntual de μ. Notar que todos los intervalos son del mismo ancho, ya que este último sólo depende de $z_{1-\alpha/2}$ una vez que se determina $\bar{x}$. En la figura 9.1b los intervalos correspondientes a las medias $\bar{x}_2$, $\bar{x}_4$ no contienen al parámetro μ, mientras que el resto de los intervalos si contienen al parámetro.

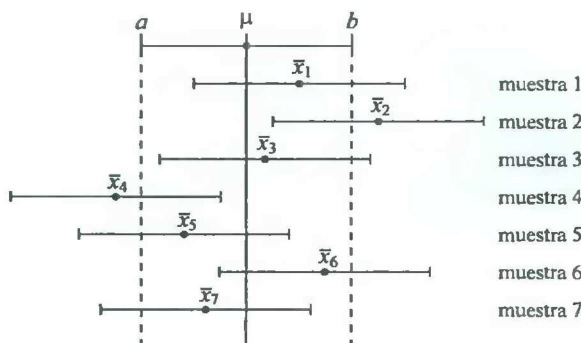


Fig. 9.1b

NOTA. (Población finita, muestreo sin reemplazo)

Si la muestra aleatoria de tamaño n es escogida *sin reposición* de una ***población finita de tamaño N***, entonces, si $n \geq 30$, la variable aleatoria:

$$Z' = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}},$$

tiene distribución aproximadamente normal $N(0,1)$. Utilizando la distribución de Z' se determina el intervalo de confianza de μ.

Luego, si $\bar{x}$ es un valor de la media $\overline{X}$ para una muestra aleatoria de tamaño n escogida de una **población *finita de tamaño N*** con varianza σ^2 supuesta conocida, el intervalo de confianza del $(1-\alpha)\times 100\%$ para μ es:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

NOTA. (Error estándar)

Se denomina **error estándar** de un estimador a la desviación estándar del estimador. A su valor numérico se denomina **error estándar estimado**. Por ejemplo, el **error estándar** (E.S.) de la media de una muestra de una población infinita (o población finita con sustitución) es

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Si la población es finita de tamaño N y el muestreo es sin reposición **el error estándar** (E.S.) de la media muestral es:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Luego, el intervalo de confianza del $(1-\alpha)\times 100\%$ para μ se puede obtener a partir de los **límites de tolerancia o confianza**:

$$\overline{x} \mp z_{1-\alpha/2} E.S.$$

EJEMPLO 9.7.

Una muestra aleatoria de 100 hogares de una ciudad indica que el promedio de los ingresos mensuales es de \$500. Encuentre un intervalo de confianza del 95% para la media poblacional de los ingresos de todos los hogares de esa ciudad. Suponga $\sigma = \$100$.

SOLUCION.

Sea X el ingreso familiar mensual de esa ciudad, cuyo promedio μ se quiere estimar a partir de una muestra aleatoria de tamaño $n = 100$. La estimación puntual de μ es $\overline{x} = 500$. Para el nivel de confianza $1-\alpha = 0.95$, en la tabla normal estándar se encuentra: $z_{1-\alpha/2} = z_{0.975} = 1.96$.

El error estándar de la media $\overline{X}$ es $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = 10.$

Los límites de tolerancia de μ son:

$$\overline{x} \mp z_{1-\alpha/2}\sigma_{\overline{X}} = 500 \mp 1.96(10) = 500 \mp 19.6$$

Luego, el intervalo de confianza del 95% para μ es: [480.4, 519.6]

Esto es, **se tiene una confianza del 95% que el promedio del ingreso familiar μ de esa ciudad, está en el intervalo [\$480.4, \$519.6]**.

NOTA. Muestras diferentes darán diferentes valores de $\overline{x}$, y por tanto darán diferentes intervalos de estimación de μ. Decir que el intervalo de estimación contiene al parámetro con confianza 95%, equivale a decir que 95 por 100 de los intervalos contienen a la media μ y que sólo el 5 por 100 no lo contienen.

EJEMPLO 9.8.

Un analista de investigación de mercados escoge una muestra aleatoria de 100 clientes de un conjunto de 500 clientes de una gran tienda que declararan ingresos mayores a \$5,000. El encuentra que los clientes de la muestra gastaron en la tienda

un promedio de $2500. Si con este valor de la muestra se estima que el gasto promedio de la población finita varía de 2446 a 2554, ¿qué nivel de confianza se utilizó?. Suponga que la desviación estándar de la población es $\sigma = \$300$.

SOLUCION.

El intervalo de confianza del $1-\alpha$ en % para la media μ, es la expresión:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

donde, $\bar{x} = \$2{,}500$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{300}{\sqrt{100}}\sqrt{\frac{500-100}{500-1}} = 26.8597$

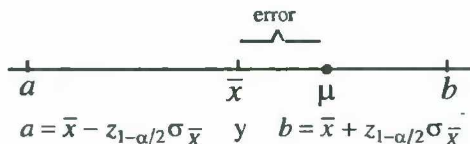
De $\mu \in [2446,\ 2554]$, se obtiene $2554 = 2{,}500 + z_{1-\alpha/2}(26.8597)$.

Luego, $z_{1-\alpha/2} = 2.01$, $\alpha = 0.0444$,

$$1-\alpha = 0.9556.$$

NOTA. (Error de estimación)

Si $\bar{x}$ estima a μ, entonces, el **error de la estimación** es el valor numérico $|\bar{x} - \mu|$ (ver figura que sigue).



El valor mínimo del error de estimación es igual a cero, esto ocurre, cuando $\bar{x}$ estima exactamente a μ.

El valor máximo del error de estimación es igual a $z_{1-\alpha/2}\sigma_{\bar{X}}$, ya que del intervalo de estimación de μ resulta:

$$|\bar{x} - \mu| \leq z_{1-\alpha/2}\sigma_{\bar{X}}.$$

Luego,

Si $\bar{x}$ estima a μ, entonces, se tiene una confianza del $(1-\alpha)\times 100\%$ de que el error de la estimación no será superior a $z_{1-\alpha/2}\sigma_{\bar{X}}$, donde $\sigma_{\bar{X}}$ es el error estándar de la media.

Por **ejemplo**, en el ejemplo 9.7 se tiene una confianza del 95% de que al estimar μ por \$500 el error de la estimación no será superior a \$19.6.

Mientras que en el ejemplo 9,8 se tiene una confianza del 95.56% de que al estimar μ como \$2500 el error de la estimación no será superior a \$54.

NOTA. (Tamaño de la muestra).

Se puede determinar que tan grande debe ser el tamaño de la muestra, n, de manera que si μ se estima por $\overline{x}$, el error de estimación no sea mayor que un valor dado e. En efecto, el valor de n se obtiene de

$$z_{1-\alpha/2}\sigma_{\overline{X}} \leq e$$

Entonces,
Si $\overline{x}$ estima a μ, entonces, **se tiene una confianza del** $(1-\alpha)\times 100\%$ **de que el error no será mayor que el valor dado** $\boldsymbol{e}$ cuando el tamaño de la muestra sea

$$n = \frac{(z_{1-\alpha/2}\sigma)^2}{e^2}$$

Si la población es finita de tamaño N y el muestreo es sin sustitución, el error estándar es $\sigma_{\overline{X}} = \left(\sigma/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$ y el valor de n se calcula por:

$$n = \frac{z_{1-\alpha/2}^2\sigma^2 N}{z_{1-\alpha/2}^2\sigma^2 + e^2(N-1)}.$$

Por **ejemplo**:

a) En el ejemplo 9.7, se tiene una confianza del 95% de que al estimar la media de la población, el error de la estimación no será mayor de \$18 cuando el tamaño n de la muestra es:

$$n = \frac{(z_{1-\alpha/2}\sigma)^2}{e^2} = \frac{(1.96)^2(100)^2}{(18)^2} = 118.5679 \cong 119.$$

b) En el ejemplo 9.8, se tiene una confianza del 97% de que al estimar la media de la población, el error de la estimación no será mayor de \$50 cuando el tamaño n de la muestra es:

$$n = \frac{z_{1-\alpha/2}^2\sigma^2 N}{z_{1-\alpha/2}^2\sigma^2 + e^2(N-1)} = \frac{(2.17)^2(300)^2(500)}{(2.17)^2(300)^2 + (50)^2(500-1)} = 126.79 \cong 127.$$

NOTA. (Estimación del total de la población)

Si la muestra aleatoria de tamaño n, se escoge de una población finita de tamaño N, entonces,

$$\textit{total de la población}: \sum_{i=1}^{N} X_i = N\mu .$$

La estimación puntual del total $N\mu$ es $N\bar{x}$.

El intervalo de confianza del ($(1-\alpha)100\%$) para μ es:

$$\bar{x} - z_{1-\alpha/2}\sigma_{\bar{X}} \le \mu \le \bar{x} + z_{1-\alpha/2}\sigma_{\bar{X}}$$

Luego, el intervalo de confianza del ($(1-\alpha)100\%$) para $N\mu$ es:

$$N(\bar{x} - z_{1-\alpha/2}\sigma_{\bar{X}}) \le N\mu \le N(\bar{x} + z_{1-\alpha/2}\sigma_{\bar{X}}).$$

donde $\sigma_{\bar{X}} = \left(\sigma/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$ es el error estándar.

Para dar un **ejemplo**, en el ejemplo 9.8, la estimación puntual del total de gastos de la población $N\mu$ es $N\bar{x} = 500(\$2500) = \$1,250,000$.

Además, los límites de confianza al 95% para el total $N\mu$ son:

$$N(\bar{x} \mp z_{1-\alpha/2}\sigma_{\bar{X}}) = 500(2500 \mp 1.96 \times 26.8597) = 1,250,000 \mp 26322.506$$

Luego, $N\mu \in [1,223,677.494,\ 1,276,322.506]$ con confianza 95%.

Consecuentemente, si el total de la población $N\mu$ se estima en \$1,250,000, se tiene una confianza del 95% de que el error de la estimación no será superior a \$26322.506.

9.4.2 Intervalo de confianza para la media μ: Varianza σ^2 supuesta desconocida

A) Población no normal

Si la población no es normal pero el tamaño de la muestra es suficientemente grande ($n \ge 30$), se utiliza la desviación estándar $\hat{s}$ de la muestra, como estimación puntual de la desviación estándar σ de la población. Entonces, utilizando

el teorema central del límite, se concluye que el intervalo de confianza del $(1-\alpha)\times 100\%$ para μ es *aproximadamente*:

$$\bar{x} - z_{1-\alpha/2}\sigma_{\bar{X}} \le \mu \le \bar{x} + z_{1-\alpha/2}\sigma_{\bar{X}}.$$

donde, el error estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ se sustituye por el error estándar estimado $\hat{\sigma}_{\bar{X}} = \hat{s}/\sqrt{n}$ si el muestreo es con o sin sustitución en una población infinita (con sustitución en una población finita de tamaño N), y se sustituye por $\hat{\sigma}_{\bar{X}} = \left(\hat{s}/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$ si el muestreo es sin sustitución en una población finita de tamaño N.

B) Población normal

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n escogida de una población normal $N(\mu, \sigma^2)$ donde la varianza σ^2 es supuesta desconocida y sean la media y la varianza muestrales respectivas:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad \hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Se sabe que las variables aleatorias:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \qquad V = \frac{(n-1)\hat{S}^2}{\sigma^2},$$

tienen distribuciones respectivas, normal $N(0,1)$ y chi-cuadrado con $n-1$ grados de libertad . Además Z y V son variables aleatorias independientes. Entonces, la variable aleatoria:

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}} = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}},$$

tiene distribución t-student con $n-1$ grados de libertad, esto es, $T \sim t(n-1)$.

Por tanto, dado el número $1-\alpha$, en la distribución de probabilidad de T se encuentran los números $\pm t_{1-\alpha/2,\,n-1}$ (figura 9.2) tales que:

$$P[-t_{1-\alpha/2,\,n-1} \le T \le t_{1-\alpha/2,\,n-1}] = 1-\alpha.$$

Al sustituir la expresión de T se obtiene:

$$P\left[-t_{1-\alpha/2,\,n-1} \le \frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}} \le t_{1-\alpha/2,\,n-1}\right] = 1-\alpha.$$

$$P[\overline{X} - t_{1-\alpha/2,\,n-1}\,\hat{S}/\sqrt{n} \le \mu \le \overline{X} + t_{1-\alpha/2,\,n-1}\,\hat{S}/\sqrt{n}] = 1-\alpha.$$

Luego,

Si $\bar{x}$ y $\hat{s}$ son la media y la desviación estándar respectivamente para un valor particular $x_1, x_2, ..., x_n$ de la muestra aleatoria de tamaño n escogida de la **población normal con varianza** σ^2 **desconocida**, entonces, el intervalo de confianza de $(1-\alpha)\times 100\%$ para μ es

$$\bar{x} - t_{1-\alpha/2,\,n-1}\,\hat{s}/\sqrt{n} \le \mu \le \bar{x} + t_{1-\alpha/2,\,n-1}\,\hat{s}/\sqrt{n}$$

El valor $t_{1-\alpha/2,\,n-1}$ se encuentra en la tabla t-student con $n-1$ grados de libertad tal que $P[T \le t_{1-\alpha/2,\,n-1}] = 1-\alpha/2$.

La ilustración es la figura 9.2, donde,

$$a = \bar{x} - t_{1-\alpha/2,\,n-1}\,\hat{s}/\sqrt{n} \quad \text{y} \quad b = \bar{x} + t_{1-\alpha/2,\,n-1}\,\hat{s}/\sqrt{n}$$

son los límites de confianza de μ inferior y superior respectivamente.

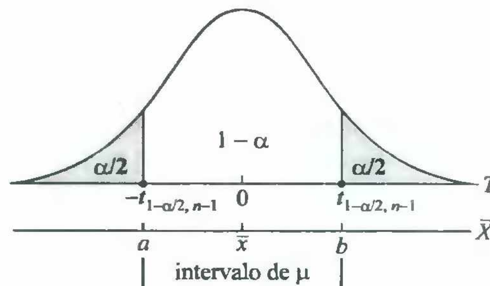


Figura 9.2: Intervalo de estimación para μ (con estadística t).

EJEMPLO 9.9.

Los contenidos de una muestra aleatoria de 5 latas de café instantáneo de un productor han dado los siguientes pesos netos en gramos:

280, 290, 285, 275, 284.

a) Encuentre un intervalo de confianza del 95% para la media de los contenidos de todas las latas de café del productor.
b) ¿Con que grado de confianza se estima que el contenido promedio de café tenga los límites de confianza 277.432 y 288.168?.

Suponga una distribución normal.

SOLUCION.

a) Sea X el peso de los contenidos de café por lata., cuyo promedio μ se quiere estimar a partir de una muestra aleatoria de tamaño $n=5$. Se supone que la distribución de X es normal con desviación estándar σ no conocida.

Para $1-\alpha=0.95$ y $n-1=4$ de libertad en la tabla t-student se encuentra $t_{1-\alpha/2,n-1}=t_{0.975,4}=2.776$.

De la muestra se obtiene $\bar{x}=282.8$ y $\hat{s}=5.63$.

El error estándar de la media $\bar{X}$ es $\hat{s}/\sqrt{5}=5.63/2.236=2.518$.

Los límites de tolerancia inferior y superior para μ son:

$$\bar{x}\mp t_{1-\alpha/2,n-1}\,\hat{s}/\sqrt{n} = 282.8\mp 2.776\times 2.518 = 282.8\mp 6.99$$

Luego, el intervalo de confianza del 95% para μ es: $[275.81, 289.79]$

b) $\mu\in[277.432, 288.168]$ con confianza $1-\alpha$. La tolerancia superior es:

$$\bar{x}+t_{1-\alpha/2,n-1}\,\hat{s}/\sqrt{n} = 282.8+t_{1-\alpha/2,4}\times 2.518 = 288.168$$

de donde resulta:

$$t_{1-\alpha/2,4}=2.132\,,\ \ 1-\alpha/2=0.95\,,\ \ \alpha=0.10\ \ \text{y}\ \ 1-\alpha=0.90\,.$$

9.5 Intervalo de confianza para la varianza

Sea $X_1, X_2,..., X_n$ una muestra aleatoria de tamaño n, escogida de una población normal con varianza σ^2, parámetro desconocido.

Un estimador puntual de la varianza σ^2 es la varianza muestral

$$\hat{S}^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}$$

cuyo valor $\hat{s}^2$ es la estimación puntual de σ^2.

Para determinar el intervalo de confianza para la varianza σ^2 se puede utilizar la estadística:

$$X = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

cuya distribución es chi-cuadrado con $n-1$ grados de libertad, esto es, $X \sim \chi^2(n-1)$ para $n \geq 2$.

Dado el grado de confianza $1-\alpha$, en la distribución $\chi^2(n-1)$ se pueden encontrar los valores $\chi^2_{\alpha/2,\,n-1}$ y $\chi^2_{1-\alpha/2,\,n-1}$ tales que (figura 9.3).

$$P[\chi^2_{\alpha/2,\,n-1} \leq X \leq \chi^2_{1-\alpha/2,\,n-1}] = 1-\alpha.$$

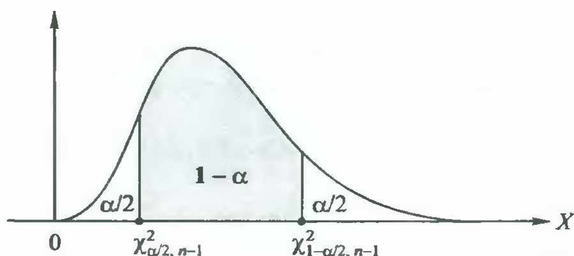


Figura. 9.3 Intervalo de confianza de la varianza σ^2.

Sustituyendo $X = (n-1)\hat{S}^2/\sigma^2$ resulta:

$$P\left[\frac{(n-1)\hat{S}^2}{\chi^2_{1-\alpha/2,\,n-1}} \leq \sigma^2 \leq \frac{(n-1)\hat{S}^2}{\chi^2_{\alpha/2,\,n-1}}\right] = 1-\alpha$$

Luego, si $\hat{s}^2$ es la varianza de una muestra aleatoria de tamaño n seleccionada de una población normal, entonces, el intervalo de confianza de $(1-\alpha)100\%$ para σ^2 es:

$$\frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha/2,\,n-1}} \leq \sigma^2 \leq \frac{(n-1)\hat{s}^2}{\chi^2_{\alpha/2,\,n-1}}$$

Los valores, $\chi^2_{\alpha/2,\,n-1}$ y $\chi^2_{1-\alpha/2,\,n-1}$ se hallan en la tabla chi-cuadrado con $n-1$ grados de libertad y con áreas acumuladas respectivas de $\alpha/2$ y $1-\alpha/2$.

EJEMPLO 9.10.

Una máquina produce piezas metálicas en forma cilíndrica. Para estimar la variabilidad de los diámetros, se toma una muestra aleatoria de 10 piezas producidas por la máquina encontrando los siguientes diámetros en centímetros:

10.1, 9.7, 10.3, 10.4, 9.9, 9.8, 9.9, 10.1, 10.3, 9.9.

Encuentre un intervalo de confianza del 95% para la varianza de los diámetros de todas las piezas producidos por la máquina. Suponga que los diámetros de las piezas se distribuyen según la normal.

SOLUCION.

Con $\alpha = 0.05$, $n = 10$ y $r = n - 1 = 9$ grados de libertad, en la tabla chi-cuadrado se encuentran:

$$\chi^2_{\alpha/2,\,n-1} = \chi^2_{0.025,9} = 2.70 \quad \text{y} \quad \chi^2_{1-\alpha/2,\,n-1} = \chi^2_{0.975,9} = 19.02\,.$$

De los datos de la muestra resulta $\hat{s}^2 = 0.056$. Los límites de confianza inferior y superior del 95% para la varianza σ^2 son respectivamente:

$$\frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha/2,\,n-1}} = \frac{9(0.056)}{19.02} = 0.0265, \qquad \frac{(n-1)\hat{s}^2}{\chi^2_{\alpha/2,\,n-1}} = \frac{(9)(0.056)}{2.70} = 0.1867.$$

Por lo tanto, el intervalo de confianza del 95% para la varianza σ^2 es:

$$0.0265 \le \sigma^2 \le 0.1867.$$

Observar que, **el intervalo de confianza del 95% para la desviación estándar** σ es:

$$0.1628 \le \sigma \le 0.432.$$

9.6 Intervalo de confianza para la razón de dos varianzas

Sean $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de dos muestras aleatorias *independientes* de tamaños n_1 y n_2 seleccionadas de dos poblaciones normales respectivas con varianzas σ_1^2 y σ_2^2.

Un estimador puntual de la razón de las varianzas σ_1^2/σ_2^2 es la estadística $\hat{S}_1^2/\hat{S}_2^2$.

Para determinar el intervalo de confianza de σ_1^2/σ_2^2 se puede utilizar la estadística F definida por:

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2},$$

que tiene distribución de probabilidad F con grados de libertad $r_1 = n_1 - 1$ y $r_2 = n_2 - 1$. Esto es, $F \sim F(r_1, r_2)$. En efecto, las variables aleatorias:

$$U = \frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2}, \text{ y, } V = \frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2}$$

tienen distribuciones respectivas chi-cuadrado con $n_1 - 1$ y $n_2 - 1$ grados de libertad.

Entonces, la variable aleatoria:

$$F = \frac{U/(n_1 - 1)}{U/(n_2 - 1)} = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene distribución F con grados de libertad $r_1 = n_1 - 1$ y $r_2 = n_2 - 1$.

Observar que para obtener tal estadística F, no se requiere asumir que las dos poblaciones tengan igual promedio.

Dado el grado de confianza $1 - \alpha$, en la distribución $F \sim F(r_1, r_2)$ se pueden encontrar los valores $f_{\alpha/2, r_1, r_2}$ y $f_{1-\alpha/2, r_1, r_2}$ (figura 9.4) tales que:

$$P[f_{\alpha/2, r_1, r_2} \le F \le f_{1-\alpha/2, r_1, r_2}] = 1 - \alpha.$$

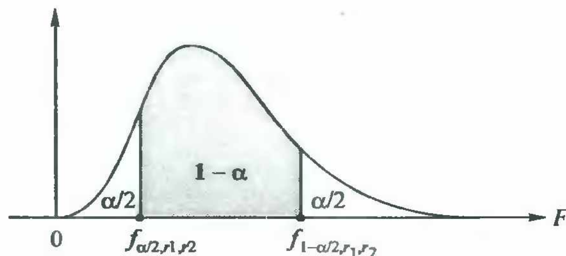


Figura. 9.4. Intervalo de confianza de la varianza σ_1^2/σ_2^2

Sustituyendo $F = (\hat{S}_1^2/\sigma_1^2)/(\hat{S}_2^2/\sigma_2^2)$ y dado que:

$$f_{\alpha/2, r_2, r_1} = \frac{1}{f_{1-\alpha/2, r_1, r_2}} \quad \text{y} \quad f_{1-\alpha/2, r_2, r_1} = \frac{1}{f_{\alpha/2, r_1, r_2}}$$

resulta.

$$P\left[\frac{\hat{S}_1^2}{\hat{S}_2^2} f_{\alpha/2, r_2, r_1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\hat{S}_1^2}{\hat{S}_2^2} f_{1-\alpha/2, r_2, r_1}\right] = 1 - \alpha.$$

Luego,

Si $\hat{s}_1^2$ y $\hat{s}_2^2$ son las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones normales, entonces, el intervalo de confianza de $(1-\alpha)\times 100\%$ para σ_1^2/σ_2^2 es:

$$\frac{\hat{s}_1^2}{\hat{s}_2^2} f_{\alpha/2, r_2, r_1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\hat{s}_1^2}{\hat{s}_2^2} f_{1-\alpha/2, r_2, r_1}$$

EJEMPLO 9.11.

Se quiere comparar la variabilidad de todas las ventas mensuales de una compañía A con la variabilidad de su competidora la compañía B. Se sabe que todas las ventas de A y de B se distribuyen normalmente. Se han tomado dos muestras aleatorias de ventas; una de 8 meses de A y otra de 6 meses de B obteniéndose las siguientes ventas:

Muestra de A: 17, 23, 21, 18, 22, 20, 21, 19.
Muestra de B: 13, 16, 14, 12, 15, 14.

Mediante un intervalo de confianza del 95% para σ_1^2/σ_2^2. ¿Se puede concluir que son iguales las varianzas de todas las ventas de las compañías A y B?.

SOLUCION.

Sean X, Y las variables aleatorias que representan las ventas de A y de B respectivamente. Se supone que las distribuciones de X, Y son normales.

Con $\alpha = 0.05$, y grados de libertad $r_1 = n_1 - 1 = 7$ y $r_2 = n_2 - 1 = 5$ en la tabla F se encuentran:

$$f_{1-\alpha/2, r_1, r_2} = f_{0.975, 7, 5} = 6.85, \text{ y, } f_{1-\alpha/2, r_2, r_1} = f_{0.975, 5, 7} = 5.29,$$

entonces, $f_{\alpha/2, r_2, r_1} = 1/f_{1-\alpha/2, r_1, r_2} = 1/f_{0.975, 7, 5} = 1/6.85 = 0.146$.

De los datos de la muestra resultan $\hat{s}_1^2 = 4.13$ y $\hat{s}_2^2 = 2$.

Los límites de confianza del 95% para σ_1^2/σ_2^2 inferior y superior son respectivamente:

$$\frac{\hat{s}_1^2}{\hat{s}_2^2} f_{\alpha/2, r_2, r_1} = \frac{4.13}{2}(0.146) = 0.30149.$$

$$\frac{\hat{s}_1^2}{\hat{s}_2^2} f_{1-\alpha/2, r_2, r_1} = \frac{4.13}{2}(5.29) = 10.9239.$$

Por lo tanto, el intervalo de confianza del 95% para la varianza σ_1^2/σ_2^2 es:

$$0.30149 \leq \sigma_1^2/\sigma_2^2 \leq 10.9239.$$

Dado que el cociente $\sigma_1^2/\sigma_2^2 = 1 \in [0.30149,\ 10.9239]$, se concluye que no hay diferencias significativas entre las varianzas de todas las ventas de A y B.

9.7 Intervalo de confianza para la diferencia entre dos medias

9.7.1 Intervalo de confianza para la diferencia entre dos medias: Varianzas σ_1^2 y σ_2^2 supuestas conocidas

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 supuestas conocidas.

Un estimador puntual de la diferencia de medias $\mu_1 - \mu_2$ es la estadística $\overline{X}_1 - \overline{X}_2$ cuyo valor $\bar{x}_1 - \bar{x}_2$ es la estimación puntual.

Si las *dos poblaciones son normales,* entonces, $\overline{X}_1$ y $\overline{X}_2$ tienen distribuciones respectivas normal $N(\mu_1, \sigma_1^2/n_1)$ y $N(\mu_2, \sigma_2^2/n_2)$ (para $n_1 \geq 2$, y $n_2 \geq 2$). En consecuencia, la estadística $\overline{X}_1 - \overline{X}_2$ tiene distribución normal $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Si las dos *poblaciones no* son *normales* pero n_1 y n_2 son suficientemente grandes ($n_1 \geq 30$ y $n_2 \geq 30$), entonces, la estadística $\overline{X}_1 - \overline{X}_2$ es aproximadamente normal $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Por tanto, según sea el caso, la variable aleatoria:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

tiene distribución exactamente o aproximadamente normal $N(0.1)$.

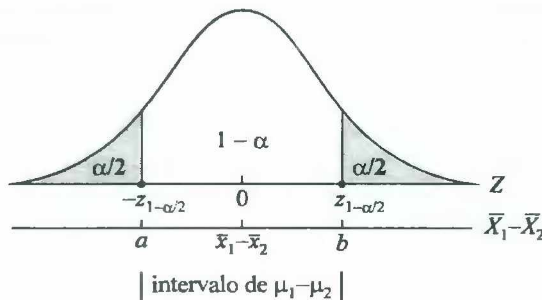


Figura 9.5: Intervalo de estimación de $\mu_1 - \mu_2$

Dado el grado de confianza $1-\alpha$, en la distribución de Z se puede encontrar el valor $z_0 = z_{1-\alpha/2}$ tal que $P[-z_0 \leq Z \leq z_0] = 1-\alpha$. (fig. 9.5)

Sustituyendo $Z = (\overline{X}_1 - \overline{X}_2)/\sigma_{\overline{X}_1-\overline{X}_2}$ donde $\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ y operando, resulta,

$$P\left[(\overline{X}_1 - \overline{X}_2) - z_0\sigma_{\overline{X}_1-\overline{X}_2} \leq \mu_1 - \mu_2 \leq (\overline{X}_1 - \overline{X}_2) + z_0\sigma_{\overline{X}_1-\overline{X}_2}\right] = 1-\alpha.$$

Luego,

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias que resultan de dos muestras independientes de tamaños n_1 y n_2 escogidas respectivamente de dos poblaciones con varianzas σ_1^2 y σ_2^2 supuestas conocidas, entonces, el intervalo de confianza del $(1-\alpha)100\%$ de $\mu_1 - \mu_2$ es:

$$(\bar{x}_1 - \bar{x}_2) - z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

El valor $z_0 = z_{1-\alpha/2}$ se obtiene de la tabla normal $N(0,1)$ de manera que $P[Z \le z_{1-\alpha/2}] = 1-\alpha/2$.

La ilustración es la figura 9.5, donde,

$$a = (\bar{x}_1 - \bar{x}_2) - z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \quad \text{y} \quad b = (\bar{x}_1 - \bar{x}_2) + z_0\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

son los límites de confianza de $\mu_1 - \mu_2$, inferior y superior respectivamente.

EJEMPLO 9.12.

Un agente de compras de una compañía está tratando de decidir si comparar la marca A o la maraca B de cierto tipo de focos ahorradores de energía. Para estimar la diferencia entre las dos marcas se lleva a cabo un experimento con dos muestras aleatorias independientes de 10 focos de cada marca resultando las medias de vida útil respectivas de 1,230 horas y 1,190 horas Estimar la verdadera diferencia de las dos medias de vida útil, mediante un intervalo de confianza del 95%. ¿Es acertada la decisión del agente si adquiere cualquiera de las dos marcas?

Suponga que las dos poblaciones tienen distribución normal con desviaciones estándares respectivas de 120 y 60 horas.

SOLUCION.

La estimación puntual de $\mu_1 - \mu_2$ es la diferencia de las medias muestrales

$$\bar{x}_1 - \bar{x}_2 = 1230 - 1190 = 40.$$

Error estándar : $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} = \sqrt{\dfrac{(120)^2}{10} + \dfrac{(60)^2}{10}} = 42.43.$

Para el grado de confianza del 95% se encuentra $z_0 = z_{1-\alpha/2} = z_{0.975} = 1.96$

Los límites de confianza inferior y superior respectivamente de $\mu_1 - \mu_2$ son

$$(\bar{x}_1 - \bar{x}_2) \mp z_0 \sigma_{\bar{x}_1 - \bar{x}_2} = 40 \mp 1.96(42.43) = 40 \mp 83.1628$$

Luego, el intervalo de confianza aproximado del 95% para $\mu_1 - \mu_2$ es

$$-43.16 < \mu_1 - \mu_2 < 123.16 .$$

Dado que $\mu_1 - \mu_2 = 0 \in [-43.16,\ 123.16]$, se concluye que $\mu_1 = \mu_2$ y que no hay diferencias significativas entre las medias de las vidas útiles de los objetos de marcas A y B. Por tanto, el agente de compras puede adquirir cualquiera de las dos.

9.7.2 Intervalo de confianza para la diferencia entre dos medias: Varianzas σ_1^2 y σ_2^2 supuestas desconocidas

A) Poblaciones no normales

Si $\bar{x}_1$ y $\bar{x}_2$ son los valores de las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones cuyas ***distribuciones son no normales*** con varianzas σ_1^2 y σ_2^2 supuestas desconocidas, entonces, siempre que los tamaños de las muestras sean grandes ($n_1 \geq 30$ y $n_2 \geq 30$), los parámetros σ_1 y σ_2 se estiman puntualmente por $\hat{s}_1$ y $\hat{s}_2$. El intervalo de confianza del $(1-\alpha)\times 100\%$ para $\mu_1 - \mu_2$ es entonces:

$$(\bar{x}_1 - \bar{x}_2) - z_0\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_0\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}\,.$$

B) Poblaciones normales

Sean $\bar{X}_1$ y $\bar{X}_2$ las medias y $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de dos muestras aleatorias independientes de tamaños n_1 y n_2 respectivamente seleccionadas de *dos poblaciones normales* con varianzas σ_1^2 y σ_2^2 supuestas desconocidas.

B1) Varianzas supuestas iguales: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

En este caso, las variables aleatorias,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}, \quad \text{y,} \quad V = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{\sigma^2},$$

tienen respectivamente distribución normal $N(0,1)$, y chi-cuadrado con grados de libertad: $n_1 + n_2 - 2$. Además, se comprueba que ambas U y V son independientes. Entonces, la variable aleatoria:

$$T = \frac{U}{\sqrt{\dfrac{V}{n_1 + n_2 - 2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\hat{S}_c^2}{n_1} + \dfrac{\hat{S}_c^2}{n_2}}}$$

tiene distribución t-student con $n_1 + n_2 - 2$ grados de libertad.

La varianza común $\hat{S}_c^2$ definida por:

$$\hat{S}_c^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

es un *estimador insesgado de la varianza común* σ^2.

Dado el grado de confianza $1-\alpha$ en la distribución $t(n_1 + n_2 - 2)$ se puede encontrar el valor $t_0 = t_{1-\alpha/2, n_1+n_2-2}$ (figura 9.6), tal que

$$P[-t_0 \le T \le t_0] = 1 - \alpha.$$

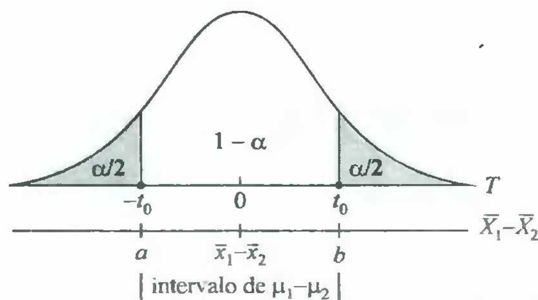


Figura 9.6: Intervalo de estimación de $\mu_1 - \mu_2$

Sustituyendo la expresión de T, y operando resulta,

$$P\left[(\bar{X}_1 - \bar{X}_2) - t_0\sqrt{\hat{S}_c^2/n_1 + \hat{S}_c^2/n_2} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + t_0\sqrt{\hat{S}_c^2/n_1 + \hat{S}_c^2/n_2}\right] = 1 - \alpha.$$

Luego,

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias que resultan de dos muestras independientes de tamaños n_1 y n_2 escogidas respectivamente de dos poblaciones normales con varianzas σ_1^2 y σ_2^2 supuestas desconocidas e iguales, entonces, el intervalo de confianza del $(1-\alpha)\times 100\%$ de $\mu_1 - \mu_2$ es:

$$(\bar{x}_1 - \bar{x}_2) - t_0\sqrt{\hat{s}_c^2/n_1 + \hat{s}_c^2/n_2} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_0\sqrt{\hat{s}_c^2/n_1 + \hat{s}_c^2/n_2}$$

El valor $t_0 = t_{1-\alpha/2, n_1+n_2-2}$, se encuentra en la tabla t-student con $n_1 + n_2 - 2$ grados de libertad tal que $P[T \le t_{1-\alpha/2, n_1+n_2-2}] = 1 - \alpha/2$.

La ilustración es la figura 9.6, donde:

$$a=(\bar{x}_1-\bar{x}_2)-t_0\sqrt{\hat{s}_c^2/n_1+\hat{s}_c^2/n_2} \quad \text{y} \quad b=(\bar{x}_1-\bar{x}_2)+t_0\sqrt{\hat{s}_c^2/n_1+\hat{s}_c^2/n_2}$$

son los límites de confianza de $\mu_1-\mu_2$, inferior y superior respectivamente.

B2) Varianzas supuestas distintas: $\sigma_1^2 \neq \sigma_2^2$

En este caso, la estadística,

$$T=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\hat{S}_1^2/n_1+\hat{S}_2^2/n_2}}$$

tiene distribución t-student con r grados de libertad, siendo,

$$r=\frac{\left[\dfrac{\hat{s}_1^2}{n_1}+\dfrac{\hat{s}_2^2}{n_2}\right]^2}{\dfrac{\left[\dfrac{\hat{s}_1^2}{n_1}\right]^2}{n_1-1}+\dfrac{\left[\dfrac{\hat{s}_2^2}{n_2}\right]^2}{n_2-1}}.$$

Dado que r rara vez es un entero, se redondea al entero más cercano.

El intervalo de confianza del $(1-\alpha)\times 100\%$ para $\mu_1-\mu_2$ es:

$$(\bar{x}_1-\bar{x}_2)-t_{1-\alpha/2,\,r}\sqrt{\hat{s}_1^2/n_1+\hat{s}_2^2/n_2}\leq \mu_1-\mu_2 \leq (\bar{x}_1-\bar{x}_2)+t_{1-\alpha/2,\,r}\sqrt{\hat{s}_1^2/n_1+\hat{s}_2^2/n_2}.$$

EJEMPLO 9.13.

El encargado de compras de una cadena de restaurantes tiene que escoger entre dos variedades de arroz A y B. Selecciona dos muestras aleatorias independientes de 10 bolsas de arroz de un kilo de cada tipo de arroz y encuentra los siguientes porcentajes de granos quebrados por kilo:

A: 6, 5, 6, 7, 4, 7, 6, 4, 3, 6.
B: 7, 6, 7, 9, 5, 8, 7, 6, 10, 8.

Estimar mediante un intervalo de confianza del 95% la diferencia promedio de porcentajes de granos quebrados por kilos de arroz de las dos variedades. ¿Se

puede aceptar que no hay diferencias significativas entre las dos medias poblacionales?.

Suponga que los porcentajes de granos quebrados por kilo en cada variedad se distribuyen normalmente con la misma varianza.

SOLUCION.

Sean X_1 y X_2 las poblaciones de porcentajes de granos quebrados por kilo. Se supone que las poblaciones son normales con varianzas desconocidas supuestas iguales.

De las muestras (utilizando el paquete *MCEST*) se obtiene:

$$n_1 = 10,\ \bar{x}_1 = 5.4,\ \hat{s}_1 = 1.35, \quad n_2 = 10,\ \bar{x}_2 = 7.3,\ \hat{s}_2 = 1.49,$$

$$\hat{s}_c^2 = \frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2} = \frac{9(1.35)^2 + 9(1.49)^2}{10+10-2} = 2.0213.$$

El **error estándar** de la diferencia de medias es:

$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\hat{s}_c^2}{n_1} + \frac{\hat{s}_c^2}{n_2}} = \sqrt{\frac{2.0213}{10} + \frac{2.0213}{10}} = 0.6358.$$

Para $1-\alpha = 0.95$ y 18 grados de libertad se halla: $t_{0.975,18} = 2.101$.

Los límites de confianza inferior y superior del 95% para $\mu_2 - \mu_1$ son:

$$(\bar{x}_2 - \bar{x}_1) \mp t_{1-\alpha/2,\, n_1+n_2-2}\, \hat{\sigma}_{\bar{x}_2-\bar{x}_1} = (7.3-5.4) \mp 2.101(0.6358) = 1.9 \mp 1.336.$$

Luego, el intervalo de confianza del 95% para $\mu_2 - \mu_1$ es

$$0.564 < \mu_2 - \mu_1 < 3.236$$

Dado que $\mu_2 - \mu_1 = 0$ no pertenece al intervalo de confianza, no se debe aceptar que $\mu_1 = \mu_2$.

EJEMPLO 9.14.

Se lleva a cabo un estudio para comparar el tiempo que tardan hombres y mujeres para realizar determinada tarea. Las experiencias anteriores indican que la distribución de tiempos tanto para hombres como para mujeres es normal con varianzas diferentes. Una muestra aleatoria de 9 hombres y 8 mujeres han dado los siguientes tiempos en minutos:

Hombres: 12, 28, 10, 25, 24, 19, 22, 33, 17 .
Mujeres: 16, 20, 16, 20, 16, 17, 15, 21.

Mediante un intervalo de confianza del 95% para la verdadera diferencia de los promedios de tiempos de hombres y mujeres, ¿se puede concluir que los hombres emplean mayor tiempo que las mujeres para hacer la tarea?.

SOLUCION.

Sean X_1 y X_2 las variables aleatorias que representan los tiempos empleados por los hombres y las mujeres respectivamente.

De las muestras dadas (utilizando el paquete *MCEST*) se obtiene:

$$n_1 = 9,\ \bar{x}_1 = 21.111, \hat{s}_1 = 7.4237\ ,$$

$$n_2 = 8,\ \bar{x}_2 = 17.625,\ \hat{s}_2 = 2.326\ .$$

La diferencia de las medias muestrales es

$$\bar{x}_1 - \bar{x}_2 = 21.111 - 17.625 = 3.486.$$

El error estándar de la diferencia de medias es

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} = \sqrt{\frac{(7.42)^2}{9} + \frac{(2.33)^2}{8}} = 2.61$$

El número de grados de libertad es

$$g = \frac{\left[\hat{s}_1^2/n_2 + \hat{s}_2^2/n_2\right]^2}{\frac{\left[\hat{s}_1^2/n_1\right]^2}{n_1 - 1} + \frac{\left[\hat{s}_2^2/n_2\right]^2}{n_2 - 1}} = \frac{\left[(7.42)^2/9 + (2.33)^2/8\right]^2}{\frac{\left[(7.42)^2/9\right]^2}{9-1} + \frac{\left[(2.33)^2/8\right]^2}{8-1}} \approx 9.73 \cong 10$$

Para $1 - \alpha = 0.95$ y $r = 10$ grados de libertad se tiene $t_{0.975,\ 10} = 2.228$.

Los límites de confianza inferior y superior aproximados de $\mu_1 - \mu_2$ son

$$\left(\bar{x}_1 - \bar{x}_2\right) \mp t_{1-\alpha/2,\ r}\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = 3.486 \mp 2.228 \times 2.61 = 3.486 \mp 5.815.$$

Luego, el intervalo de confianza del 95 por 100 para $\mu_1 - \mu_2$ es

$$-2.329 < \mu_1 - \mu_2 < 9.301$$

Dado que $\mu_1 - \mu_2 = 0 \in [-2.329,\ 9.301]$ podemos concluir que $\mu_1 = \mu_2$, por tanto, los tiempos promedios de hombres y mujeres son iguales.

Se recomienda al lector resolver este problema usando un **paquete de computo**, por ejemplo el *MCEST*.

9.8 Intervalo de confianza para la diferencia entre dos medias con observaciones pareadas

Sea (X_1,Y_1), (X_2,Y_2),..., (X_n,Y_n) una muestra aleatoria de n datos aparejados, donde las muestras $X_1,X_2,...,X_n$, e $Y_1,Y_2,...,Y_n$, **correlacionadas**, son seleccionadas respectivamente de dos poblaciones normales $X \sim N(\mu_1,\sigma_1^2)$ y $Y \sim N(\mu_2,\sigma_2^2)$.

Podemos concebir estas n diferencias: $D_1 = X_1 - Y_1$, $D_2 = X_2 - Y_2$,..., $D_n = X_n - Y_n$ como una muestra aleatoria seleccionada de una población de diferencias $D = X - Y$ cuya distribución es normal $N(\mu_D,\sigma_D^2)$, con media $\mu_D = \mu_1 - \mu_2$ y varianza $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\text{cov}(X,Y)$.

El estimador insesgado de $\mu_D = \mu_1 - \mu_2$ es la estadística:

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{\sum_{i=1}^{n}(X_i - Y_i)}{n} = \frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{X} - \overline{Y}$$

cuyo valor $\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ es la estimación insesgada de μ_D.

Si σ_D^2 es conocida, la estadística $\overline{D}$ tiene distribución normal $N(\mu_D,\sigma_D^2/n)$. Consecuentemente la estadística:

$$Z = \frac{\overline{D} - (\mu_1 - \mu_2)}{\sigma_D/\sqrt{n}}$$

tiene distribución normal $N(0,1)$. Esta estadística se utiliza para determinar el intervalo e confianza de $\mu_D = \mu_1 - \mu_2$ con varianza σ_D^2 conocida.

Luego, si $\overline{d}$ es la media de las diferencias de n datos pareados escogidos de una distribución normal con desviación estándar σ_D supuesta conocida, entonces, el

intervalo de confianza del $(1-\alpha)\times 100\%$ para la diferencia de medias $\mu_D = \mu_1 - \mu_2$ es la expresión:

$$\bar{d} - z_{1-\alpha/2}\,\sigma_D/\sqrt{n} \leq \mu_1 - \mu_2 \leq \bar{d} + z_{1-\alpha/2}\,\sigma_D/\sqrt{n}$$

siendo $z_{1-\alpha/2}$ un valor que se encuentra en la tabla normal $N((0,1)$, tal que

$$P[Z \leq z_{1-\alpha/2}] = 1-\alpha/2$$

Por otra parte, si la varianza σ_D^2 es *desconocida*, el estimador puntual de esta varianza, es la estadística:

$$\hat{S}_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$$

cuyo valor,

$$\hat{s}_d^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} = \frac{\sum_{i=1}^{n} d_i^2 - n(\bar{d})^2}{n-1}$$

es la estimación de σ_D^2. Además se verifica que la estadística

$$T = \frac{\bar{D} - (\mu_1 - \mu_2)}{\hat{S}_D/\sqrt{n}}$$

tiene distribución t-student con $n-1$ grados de libertad. Esta estadística se usa para determinar el intervalo de confianza de $\mu_D = \mu_1 - \mu_2$

Luego,

Si $\bar{d}$ y $\hat{s}_d$ son la media y la desviación estándar de una muestra aleatoria de n diferencias de pares de datos de una población normal con varianza σ_D^2 *supuesta desconocida*, entonces el intervalo de confianza del $(1-\alpha)\times 100\%$ para $\mu_D = \mu_1 - \mu_2$ es

$$\bar{d} - t_{1-\alpha/2,n-1}\,\hat{s}_d/\sqrt{n} \leq \mu_1 - \mu_2 \leq \bar{d} + t_{1-\alpha/2,n-1}\,\hat{s}_d/\sqrt{n}$$

El valor $t_{1-\alpha/2,n-1}$ se encuentra en la tabla t-student con $n-1$ grados de libertad, tal que: $P[T \leq t_{1-\alpha/2,n-1}] = 1-\alpha/2$.

EJEMPLO 9.15.

En estudios generales ciencias de la PUCP se han escogido 12 pares de alumnos sobre la base de la similaridad se sus rendimientos. A un alumno de cada par le fue enseñado el curso de cálculo I por el método tradicional (X) y al otro alumno por el método de talleres (Y). Estos alumnos dieron una prueba con los siguientes resultados:

Par de Alumnos	1	2	3	4	5	6	7	8	9	10	11	12
Tradicional (X_i)	14	15	12	13	15	11	10	15	15	16	14	8
Talleres (Y_i)	12	16	12	11	12	09	07	13	14	15	12	10
Dfirerencias d_i	2	−1	0	2	3	2	3	2	1	1	2	−2
d_i^2	4	1	0	4	9	4	9	4	1	1	4	4

Mediante un intervalo de confianza del 95%, determinar si los dos métodos son igualmente efectivos. Suponga distribución normal.

SOLUCION.

Haciendo $d_i = X_i - Y_i, \quad i = 1, 2, \cdots 12$, se obtiene,

$$\bar{d} = \frac{\sum_{i=1}^{12} d_i}{12} = \frac{15}{12} = 1.25.$$

$$\hat{s}_d = \sqrt{\frac{\sum_{i=1}^{12}(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{45 - 12(1.25)^2}{12-1}} = 1.545$$

Error estándar: $\hat{s}_d / \sqrt{n} = 1.545 / \sqrt{12} = 0.446.$

Para $1-\alpha = 0.95$ y $n-1 = 11$ grados de libertad se determina $t_{1-\alpha/2, n-1} = t_{0.975, 11} = 2.201.$

Los límites de confianza para $\mu_D = \mu_1 - \mu_2$ son,

$$\bar{d} \mp t_{1-\alpha/2, n-1}\, \hat{s}_d / \sqrt{n} = 1.25 \mp 2..201 \times 0.446 = 1.25 \mp 0.9816$$

Por tanto, $\mu_D = \mu_1 - \mu_2 \in [0.268,\ 2.232]$ con confianza del 95%

Como $0 \notin [0.268, 0.232]$, entonces, $\mu_1 \neq \mu_2$, aún más, $\mu_1 > \mu_2$.

9.9 Intervalo de confianza para una proporción

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n escogida de una población de Bernoullí $B(1, p)$, cuyo parámetro p es la proporción de éxitos en la población. En la muestra cada $X_i = 1$, si hay éxito con probabilidad p, y cada $X_i = 0$, si no hay éxito con probabilidad $1 - p$.

El estimador puntual del parámetro p es la estadística $\overline{P}$, *proporción de éxitos en la muestra* definida por:

$$\overline{P} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{o} \quad \overline{P} = \frac{X}{n}$$

donde la variable aleatoria $X = \sum_{i=1}^{n} X_i$ es el número de éxitos en la muestra y cuya distribución es Binomial $B(n, p)$.

El valor $\overline{p} = x/n$ que se obtiene de $\overline{P}$ para una muestra especifica, es la estimación puntual del parámetro p.

La estadística $\overline{P}$ tiene:

media $\mu_{\overline{P}} = E(\overline{P}) = p$

y varianza $\sigma^2_{\overline{P}} = Var(\overline{P}) = p(1-p)/n$,

Además, si el tamaño n de la muestra, es suficientemente grande ($n \geq 30$), por el teorema central del límite, la distribución de probabilidad de la proporción muestral $\overline{P}$ es aproximadamente la normal, con media p, varianza $p(1-p)/n$.

Luego, la variable aleatoria estandarizada

$$Z = \frac{\overline{P} - p}{\sqrt{p(1-p)/n}}$$

tiene distribución aproximadamente $N(0,1)$.

Se puede usar la distribución de Z para determinar un intervalo de confianza del parámetro p.

Dado el grado de confianza $1-\alpha$, podemos encontrar en la distribución de Z el valor $z_{1-\alpha/2}$ tal que (figura 9.7) $P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1-\alpha$..

Sustituyendo la expresión de Z resulta:

$$P[-z_{1-\alpha/2} \leq \frac{\overline{P} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}] = 1 - \alpha.$$

$$P[\overline{P} - z_{1-\alpha/2}\sqrt{p(1-p)/n} \leq p \leq \overline{P} + z_{1-\alpha/2}\sqrt{p(1-p)/n}] = 1 - \alpha.$$

El error estándar $\sigma_{\overline{P}} = \sqrt{p(1-p)/n}$ se estima por $\hat{\sigma}_{\overline{P}} = \sqrt{\overline{p}(1-\overline{p})/n}$.
Luego,

Si $\overline{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, entonces, el intervalo de confianza del $(1-\alpha)\times 100\%$ para p es:

$$\overline{p} - z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \leq p \leq \overline{p} + z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

El valor $z_{1-\alpha/2}$ se halla en la tabla normal $N(0,1)$, de manera que

$$P[Z \leq z_{1-\alpha/2}] = 1 - \alpha/2.$$

La ilustración es la figura 9.7, donde

$$a = \overline{p} - z_{1-\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n} \quad \text{y} \quad b = \overline{p} + z_{1-\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n}$$

son los límites de confianza de p, inferior y superior respectivamente.

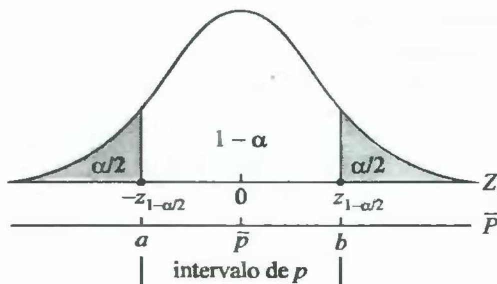


Figura 9.7: Intervalo de estimación del $(1-\alpha)\times 100\%$ para p

EJEMPLO 9.16.

Una encuestadora utilizó una muestra aleatoria de 600 electores que acaban de votar y encontró que 240 votaron a favor del candidato A.

a) Estimar el porcentaje de electores a favor de A en toda la población, utilizando un intervalo de confianza del 95%.
b) Si la proporción a favor del candidato A se estima en 40%, ¿cuánto es el error máximo de la estimación, si se quiere tener una confianza del 98%?.
c) Si con la misma muestra la proporción a favor de B se estima en 38% con una confianza del 98% que el error no es mayor a 4.62%, ¿Se puede proclamar a A como ganador de las elecciones?.
d) ¿Qué tan grande se requiere que sea la muestra si se desea tener una confianza del 94% de que el error de estimación de p no sea superior a 2%?.

SOLUCION.

a) La estimación puntual de la proporción p a favor de A en la población, es la proporción a su favor en la muestra de $n = 600$ electores; esto es, $\overline{p} = 240/600 = 0.40$.

La estimación del error estándar es:

$$\hat{\sigma}_{\overline{p}} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = \sqrt{\frac{(0.40)(0.60)}{600}} = 0.02.$$

Para $1-\alpha = 0.95$ se tiene $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Los límites de confianza de p, inferior y superior, son respectivamente

$$\overline{p} \mp z_{1-\alpha/2}\hat{\sigma}_{\overline{p}} = 0.40 \mp 1.96(0.02) = 0.40 \mp 0.0392$$

Luego, el intervalo de confianza del 95% para p es de 0.3608 a 0.4392. Es decir, $p \in [36.08\%, 43.92\%]$ con confianza del 95%.

b) Si p se estima por $\overline{p}$ se tiene una confianza del $(1-\alpha)\times 100\%$ que el error de la estimación no será mayor que $z_{1-\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n}$.

Para una confianza del 98%, $z_{1-\alpha/2} = z_{0.99} = 2.33$, y

$$z_{1-\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n} = 2.33\sqrt{(0.40)(0.60)/600} = 0.0466.$$

Luego, si con $n = 600$, p se estima en 0.40, se tiene una confianza del 98% de que el error de la estimación a favor de A no será mayor a 4.66%.

c) El intervalo de confianza del 98% a favor de A es [35.34%, 44.66%].
El intervalo de confianza del 98% a favor de B es [33.38%, 42.62%].

Dado que la intersección de los intervalos no es vacío, no se puede proclamar a A como ganador. En este caso se dice que hay un **empate técnico**.

d) Dado el error máximo e de la estimación de p con confianza de $(1-\alpha)\times 100\%$ el tamaño n de la muestra, se puede determinar en dos formas:

d1) Si se tiene el valor de $\overline{p}$ (de una *muestra preliminar o piloto*), el error máximo de estimación de p es:

$$e = z_{1-\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n}$$

de donde resulta; $$n = \frac{(z_{1-\alpha/2})^2\,\overline{p}(1-\overline{p})}{e^2}.$$

En nuestro ejemplo $\overline{p} = 0.60$. Para el nivel de confianza $1-\alpha = 0.94$, se obtiene $z_{1-\alpha/2} = z_{0.97} = 1.88$. Luego, se tiene una confianza del 94% que el error al estimar p no será mayor que 0.02 si el tamaño de la muestra es,

$$n = (1.88)^2(0.6)(0.4)/(0.02)^2 = 2120.64 \cong 2121.$$

d2) Si no se tiene el valor $\overline{p}$, entonces, se puede usar el valor máximo $\overline{p} = \overline{q} = 1/2$.

En efecto, $\overline{p}\times\overline{q} = \overline{p}(1-\overline{p}) = -(\overline{p}-1/2)^2 + 1/4 \leq 1/4$

Luego, de $e = z_{1-\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n}$, resulta, $$n \cong \frac{(z_{1-\alpha/2})^2}{4e^2}.$$

Para $1-\alpha = 0.94$, $z_{1-\alpha/2} = z_{0.97} = 1.88$. Luego, se tiene una confianza del 94% que el error al estimar p no será mayor de 0.02 si el tamaño de la muestra es,

$$n = (1.88)^2/(4\times(0.02)^2) = 2209.$$

NOTA. Si el muestreo es sin reemplazo en una población (Bernoullí de valores 0 y 1) finita de tamaño N, entonces el error estándar de $\overline{P}$ es $\sigma_{\overline{P}} = \left(pq/\sqrt{n}\right)\sqrt{(N-n)/(N-1)}$. El error estándar estimado de $\overline{P}$ es:

$$\hat{\sigma}_{\overline{P}} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\sqrt{\frac{N-n}{N-1}}.$$

y el valor de n se calcula por

$$n = \frac{z_{1-\alpha/2}^2\,\overline{pq}N}{z_{1-\alpha/2}^2\,\overline{pq} + e^2(N-1)}.$$

Si se desconoce, $\overline{p}$, se puede utilizar el valor $\overline{p}$=0.5.

EJEMPLO 9.17.

Una empresa va a hacer un estudio de mercado antes de lanzar un nuevo producto hacia una población de 30,000 consumidores.

a) ¿Qué tamaño de muestra deberá escoger si quiere tener una confianza del 95% de que error de la estimación de la proporción a favor del producto no sea superior al 4%?.

b) Si con el tamaño de la muestra calculado en a) se utiliza $\overline{p}=0.7$ como estimación de la proporción de todos los consumidores que prefieren su producto. ¿Qué grado de confianza utilizó si estimó de 19,783 a 22,217 el total de consumidores de la población que prefieren su producto?

SOLUCION.

a) Para $1-\alpha=0.95$, resulta $z_{1-\alpha/2}=z_{0.975}=1.96$.

Utilizando el valor $\overline{p}(1-\overline{p})=1/4$ y $N=30{,}000$ se tiene:

$$n=\frac{z^2_{1-\alpha/2}\overline{pq}N}{z^2_{1-\alpha/2}\overline{pq}+e^2(N-1)}=\frac{(1.96)^2(30{,}000)}{(1.96)^2+4(0.04)^2(30{,}000-1)}=588.49\cong 589$$

b) El intervalo $19{,}783\le Np\le 22{,}217$, resulta de $N(\overline{p}\mp z_{1-\alpha/2}\sigma_{\overline{P}})$, donde $\sigma_{\overline{P}}=\sqrt{(\overline{p}(1-\overline{p})/n)((N-n)/(N-1))}$.

Para $n=589$, $N=30{,}000$ y $\overline{p}=0.70$, se obtiene $\sigma_{\overline{P}}=0.0187$.

De $22{,}217=N(\overline{p}+z_{1-\alpha/2}\sigma_{\overline{P}})$ resulta: $z_{1-\alpha/2}=2.17$, $1-\alpha/2=0.985$.
$\alpha=0.03$, y $1-\alpha=0.97$

9.10 Intervalo de confianza para la diferencia entre dos proporciones

Sean $\overline{P}_1$ y $\overline{P}_2$ las proporciones de éxitos de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones de Bernoullí $B(1,p_1)$ y $B(1,p_2)$, donde p_1 y p_2 son los respectivos parámetros proporciones de éxitos.

La estimación puntual de p_1-p_2 es la estadística $\overline{P}_1-\overline{P}_2$ cuyo valor es $\overline{p}_1-\overline{p}_2$.

Si n_1 y n_2 son suficientemente grandes, $\overline{P}_1$ y $\overline{P}_2$ tienen distribuciones aproximadamente normales respectivas $N(p_1, p_1(1-p_1)/n_1)$ y $N(p_2, p_2(1-p_2)/n_2)$. Por tanto, por la propiedad reproductiva de la normal la estadística $\overline{P}_1 - \overline{P}_2$ tendrá distribución aproximadamente normal con media:

$$\mu_{\overline{P}_1-\overline{P}_2} = E(\overline{P}_1 - \overline{P}_2) = p_1 - p_2,$$

y con varianza

$$\sigma^2_{\overline{P}_1-\overline{P}_2} = Var(\overline{P}_1 - \overline{P}_2) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2.$$

Por consiguiente,

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sqrt{p_1q_1/n_1 + p_2q_2/n_2}},$$

tendrá distribución aproximadamente normal $N(0,1)$, siendo $q_1 = 1 - p_1$ y $q_2 = 1 - p_2$. Esta distribución se utiliza para determinar el intervalo de confianza de $p_1 - p_2$.

Dado el grado de confianza $1-\alpha$, en la distribución de Z, se puede encontrar el valor $z_0 = z_{1-\alpha/2}$ tal que

$$P[-z_0 \le Z \le z_0] = 1 - \alpha.$$

Sustituyendo el valor de Z y luego de hacer operaciones convenientes se obtiene

$$P\left[(\overline{P}_1 - \overline{P}_2) - z_0\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \le p_1 - p_2 \le (\overline{P}_1 - \overline{P}_2) + z_0\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right] = 1 - \alpha.$$

El error estándar:
$$\sigma_{\overline{P}_1-\overline{P}_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}},$$

se estima por:
$$\hat{\sigma}_{\overline{P}_1-\overline{P}_2} = \sqrt{\frac{\overline{p}_1\overline{q}_1}{n_1} + \frac{\overline{p}_2\overline{q}_2}{n_2}},$$

Luego,

Si $\overline{p}_1$ y $\overline{p}_2$ son las proporciones de éxitos en dos muestras aleatorias independientes de tamaños n_1 y n_2 respectivamente, entonces, el intervalo de confianza del $(1-\alpha)\times 100\%$ de $p_1 - p_2$ es

$$(\overline{p}_1 - \overline{p}_2) - z_{1-\alpha/2}\hat{\sigma}_{\overline{p}_1-\overline{p}_2} \le p_1 - p_2 \le (\overline{p}_1 - \overline{p}_2) + z_{1-\alpha/2}\hat{\sigma}_{\overline{p}_1-\overline{p}_2}$$

El valor $z_{1-\alpha/2}$ se halla en la tabla normal $N(0,1)$, de manera que $P[Z \leq z_{1-\alpha/2}] = 1-\alpha/2$.

EJEMPLO 9.18.

Un fabricante afirma que su nuevo producto de consumo popular prefieren más los hombres que las mujeres. Para comprobar tal afirmación, se toma una muestra aleatoria de 250 hombres y otra de 200 mujeres, y se encuentra que 175 hombres y 120 mujeres prefieren el nuevo producto. Utilizando un intervalo de confianza del 95% para la verdadera diferencia de proporciones de preferencias entre los hombres y las mujeres, ¿se puede concluir que el fabricante del nuevo producto tiene la razón?.

SOLUCION.

De los datos del problema se obtiene:

$$\overline{p}_1 = 175/250 = 0.7\,, \quad \overline{p}_2 = 120/200 = 0.6\,,$$

$$\overline{p}_1 - \overline{p}_2 = 0.7 - 0.6 = 0.1$$

La estimación puntual del parámetro $p_1 - p_2$, es la diferencia de proporciones muestrales $\overline{p}_1 - \overline{p}_2 = 0.1$

El error estándar de la diferencia de proporciones, $\overline{P}_1 - \overline{P}_2$ es:

$$\hat{\sigma}_{\overline{P}_1-\overline{P}_2} = \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}} = \sqrt{\frac{(0.7)(0.3)}{250} + \frac{(0.6)(0.4)}{200}} = 0.045.$$

Para $1-\alpha = 0.95$, se tiene $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Los límites de tolerancia para $p_1 - p_2$ son:

$$(\overline{p}_1 - \overline{p}_2) \mp z_{1-\alpha/2}\hat{\sigma}_{\overline{p}_1-\overline{p}_2} = 0.1 \mp 1.96(0.045) = 0.1 \mp 0.0882.$$

En consecuencia, el intervalo de confianza del 95% para $p_1 - p_2$ es

$$0.0118 \leq p_1 - p_2 \leq 0.1882.$$

Dado que el intervalo resultante no contiene el valor cero, debemos concluir que las proporciones de preferencias en la población son diferentes, esto es, $p_1 \neq p_2$, y dado que el intervalo contiene valores positivos, hay razones para concluir que $p_1 > p_2$.

EJERCICIOS

Una media

1. Una máquina llena un determinado producto en bolsas cuyo peso medio es μ gramos. Suponga que la población de los pesos es normal con desviación estándar 20 gramos.
   a) Estime μ de manera que el 99.38% de las bolsas tengan pesos no superiores a 550 gramos.
   b) Estime μ mediante un intervalo de confianza del 95%, si una muestra aleatoria de 16 bolsas ha dado una media de 495 gramos

   Rp. a) 500 b) 495 ∓ 9.8.

2. Se decide estimar la media μ del nivel de ansiedad de todos los estudiantes preuniversitarios. Se supone que la población de los puntajes de la prueba para medir la ansiedad se distribuye normalmente con desviación estándar igual a 10 puntos.
   a) Determinar el intervalo para μ con confianza del 95%, si una muestra aleatoria de tamaño 100 ha dado una media de 70 puntos.
   b) Si μ se estima en 70 puntos con el nivel de confianza del 98%, ¿es el error de la estimación puntual superior a 5 puntos?
   c) Si Ud. considera que el intervalo encontrado en a) no es muy preciso, ¿qué acción debería tomar para que el intervalo de estimación al 95% sea más preciso?.

   Rp. a) 70 ∓ 1.96 b) No, es =2.33, c) aumentar el tamaño de la muestra.

3. El tiempo en minutos que utilizan los clientes en sus distintas operaciones en un banco local es una variable aleatoria cuya distribución se supone normal con una desviación estándar de 3 minutos. Se han registrado los tiempos de las operaciones de 9 clientes del banco resultando una media igual a 9 minutos:
   a) Hallar el nivel de confianza si la estimación de μ es el intervalo de 7 a 11 minutos.
   b) Si μ se estima por $\bar{x}$, calcular la probabilidad de que la media de los tiempos de todas las muestras de tamaño 9 esté entre 6.5 y 11.5 minutos.

   Rp. a) 0.9544. b) 0.9876

4. Un fabricante afirma que el peso promedio de las latas de fruta en conserva que saca al mercado es 19 onzas. Para verificar esta afirmación se escogen al azar 20 latas de la fruta y se encuentra que el peso promedio es 18.5 onzas Suponga que la población de los pesos es normal con una desviación estándar de 2 onzas.
   a) Utilizando un intervalo de confianza del 98% para μ, ¿se puede aceptar la afirmación del fabricante?

b) ¿Qué tamaño de muestra se debe escoger para estimar μ si se quiere un error no superior a 0.98 onzas con confianza del 95%?.

Rp. a) 18.5 ∓ 0.932 ,si, b) $n \cong 16$.

5. Se quiere hacer una encuesta para estimar el tiempo promedio por semana que los niños ven televisión. Por estudios anteriores se sabe que la desviación estándar de dicho tiempo es de 3 horas. Con el nivel de confianza del 99%.

a) ¿Qué tamaño de muestra se debería elegir si el error de la estimación puntual no es superior a media hora?

b) ¿Qué costo se debe presupuestar para hacer la encuesta si ésta tiene un costo fijo de $5000 más un costo variable de $2 por cada entrevista,?

Rp. a) $n \cong 238.7 \cong 239$, b) $5478.

6. Un fabricante produce focos cuya duración tiene distribución normal . Si una muestra aleatoria de 9 focos da las siguientes vidas útiles en horas

775, 780, 800, 795, 790, 785, 795, 780, 810

a) Estimar la duración media de todos los focos del fabricante mediante un intervalo de confianza del 95%.

b) Si la media poblacional se estima en 790 horas con una confianza del 98%, ¿cuánto es el error máximo de la estimación si se quiere una confianza del 98%?

Rp. Utilizando el paquete *MCEST* se tiene, a) 790 ∓ 8.59, b) 10.79.

7. Para determinar el rendimiento anual de ciertos valores, un grupo de inversionistas tomó una muestra aleatoria de 49 de tales valores encontrando una media de 8.71% y una desviación estándar $\hat{s} = 2.1\%$.

a) Estime el verdadero rendimiento anual promedio de tales valores mediante un intervalo de confianza del 96%.

b) Calcule el riesgo α si el rendimiento anual promedio de todos los valores se estima entre 7.96% y 9.46%. Rp. a) 8.71 ∓ 0.615, b) 0.0124

8. La duración de cierto tipo de batería es una variable aleatoria cuya distribución se supone normal. Inicialmente se estima que la duración media es de 500 horas y que el 95% duran entre 480.4 y 519.6 horas. Si se eligen 9 baterías al azar y se encuentra que la duración media es 480 horas. Utilizando un intervalo de confianza del 95% para la media μ, ¿se debería inferir que la duración media es diferente de 500 horas?.

Rp. $\sigma = 10$, 480 ∓ 6.53, si

9. Encontrar el tamaño de muestra que se debe tomar para estimar la media de las longitudes de los tornillos que produce una fábrica con un error no mayor de 0.0233 cm. al nivel de confianza del 98%, si; además se indica que la longitud

de los tornillos tiene distribución normal y si la longitud se desvía de la media en a lo más 0.08 cm. con probabilidad 0.9544.

Rp. $\sigma = 0.04$, $n = 16$

10. Las cajas de un cereal producidos por una fabrica deben tener un contenido promedio de 160 gramos. Un inspector de INDECOPI tomó una muestra aleatoria de 10 cajas para calcular los pesos X_i en gramos. Si de la muestra resultan las siguientes sumas:

$$\sum_{i=1}^{10} X_i = 1590, \quad \sum_{i=1}^{10} X_i^2 = 252{,}858$$

Mediante un intervalo de confianza del 98% para μ, ¿es razonable que el inspector multe al fabricante?. Suponga que el peso de las cajas del cereal tiene distribución normal.

Rp. $\bar{x} = 159$, $\hat{s} = 2.309$, ES=0.73, 159 ∓ 2.06. No.

11. El ingreso mensual de cada una de las 500 microempresas de servicios de una ciudad, es una variable aleatoria con media μ desconocida. Con el fin de simplificar la recaudación de impuestos, la Sunat ha dispuesto que a estas empresas se las grave mensualmente con un 10% de sus ingresos. De una muestra al azar de 50 microempresas se obtuvo un ingreso mensual promedio de $1000 con una desviación estándar de $80.

a) Estime el monto medio de los ingresos de las microempresas de la ciudad con un intervalo de confianza del 95%
b) Estime el monto promedio de la recaudación a estas microempresas con un intervalo de confianza del 95%
c) Si el propósito de la Sunat es lograr mensualmente una recaudación total de al menos $52,000 a estas microempresas, ¿es factible que se cumplan sus metas?, ¿por qué?.

Rp. a) IC de ingresos μ: 1000 ± 21.06. b) $R = 0.1X$, IC de μ_R: [97.894, 102.106],
c) IC del total: [48,947, 51,053], 52,000 no está en el IC. No es posible

12. Un auditor escoge una muestra aleatoria de 15 cuentas por cobrar de un total de 400 cuentas de una compañía y encuentra las siguientes cuentas en dólares

730, 759, 725, 740, 754, 745, 750, 753, 730, 780, 725, 790, 719, 775, 700

Utilizando un intervalo de confianza del 95%, estime

a) El monto promedio por cuentas por cobrar.
b) El monto total de todas las cuentas por cobrar.

Suponga que las 400 cuentas se distribuyen aproximadamente normal.

Rp. $\bar{x} = 745$, $\hat{s} = 24.6287$, población finita, ES=6.247, gl=14, a) 745 ∓ 13.4, b) $400(745 \mp 13.4)$

13. Para la campaña de Navidad una fábrica debe manufacturar 2000 juguetes de cierto tipo. Si una muestra aleatoria de 36 tiempos de fabricación en horas, $x_1, x_2, ..., x_{36}$, de tales juguetes ha dado,

$$\sum x_i = 108 \ , \ \sum x_i^2 = 325.4$$

a) Estime el tiempo promedio por juguete mediante un intervalo de confianza del 97%

b) Estime el tiempo total que se requiere para fabricar los 2000 juguetes mediante un intervalo de confianza del 97%

Rp. $\bar{x} = 3$, $\hat{s} = 0.2$, ES=0.033. a) 3 ∓ 0.072, b) 6000 ∓ 143.22

14. Un comerciante estima en \$55,000 el costo total de 3000 unidades de mercadería de diverso tipo que posee. Para verificar esta estimación va a escoger una muestra aleatoria de n unidades para hacer una estimación del costo total. Suponga que la población de los costos es normal con $\sigma = \$2.5$ por artículo. Calcular el valor de n si se requiere con confianza del 95% un error de la estimación no superior a 0.6844.

Rp. Población finita, n=50.4≅51

Una proporción

15. En un estudio socioeconómico se tomó una muestra aleatoria de 100 comerciantes informales y se encontró entre otros datos los siguientes: un ingreso medio de \$600, una desviación estándar de \$50 y sólo el 30% tienen ingresos superiores a \$800.

a) Estimar la proporción de todos los comerciantes con ingresos superiores a \$800, mediante un intervalo de confianza del 98%.

b) Si la proporción de todos los comerciantes con ingresos superiores a \$800 se estima entre 20.06% y 39.94% ¿qué grado de confianza se utilizó?

Rp. a) 0.3 ∓ 0.10677, b) $1-\alpha=0.97$.

16. Una muestra aleatoria de 400 menores de 16 años revela que 220 consumen licor.

a) Estimar la proporción de menores de 16 años que consumen licor en toda la población mediante una intervalo de confianza del 99% .

b) ¿Qué se puede afirmar con confianza del 99% acerca de la posible magnitud del error si se estima que el porcentaje de menores de 16 años que consumen alcohol es 0.55? Rp. a) 0.55 ∓ 0.064, b) 0.064

17. Un fabricante afirma que es 5% el porcentaje de piezas con algún tipo de defecto que resulta del total de la producción. Para verificar tal afirmación se toma una muestra aleatoria de 100 piezas y se encuentra que 10% ellos tienen

algún tipo de defecto. Mediante un intervalo de confianza del 95% para la proporción de piezas defectuosas de toda la producción. ¿es aceptable la afirmación del fabricante?.

Rp. 0.1 ∓ 0.0588, $0.5 \in$ I.C. se acepta la afirmación del fabricante

18. Dos candidatos A y B compiten como favoritos en las próximas elecciones. En la última encuesta a partir de una muestra grande de electores se estima con una misma confianza que A tendría 40% de los votos con un error máximo de 3%, mientras que B tendría entre 31% y 39% de los votos.

a) En base a esta encuesta , ¿cuál de los dos candidatos sería el ganador absoluto?.

b) ¿Qué tamaño de muestra se debe elegir si se quiere tener una confianza del 98% de que el error de estimación de todos los electores a favor de A no sea superior al 2%?.

Rp. a) Cualquiera, pues hay empate técnico, c) 3258 o 3394 (sin usar proporción).

19. La oficina de planificación familiar de cierta provincia quiere estimar el porcentaje de familias con más de 4 hijos.

a) ¿Qué tamaño de muestra se requiere para asegurar con una confianza del 95% que el error de la estimación de tal porcentaje no sea superior a 0.05?

b) Si en una muestra aleatoria de 385 familias se encuentra que 154 de ellas tienen más de 4 hijos, estime el porcentaje de familias con más de 4 hijos en toda la provincia, mediante un intervalo de confianza del 98%.

Rp. a) 385, b) 0.4 ∓ 0.058

20. Se desea realizar un estudio de mercado para determinar la proporción de amas de casa que prefieren una nueva pasta dental.

a) Si la encuesta tiene un costo fijo de $500 más un costo variable de $5 por cada entrevista, ¿cuánto debería costar la encuesta si se desea que el error al estimar la proporción verdadera no sea mayor que 2%, con un nivel de confianza del 97%?.

b) Si para el tamaño de muestra hallado en a) se encuentra que 736 prefieren la nueva pasta dental, estimar la proporción verdadera con un coeficiente de confianza de 99%?

Rp. a) n=2944, costo total = 2944×5$+500$=15,220$, b) 0.25 ∓ 0.01859

21. Para estimar el porcentaje de todos los electores a favor de un candidato, una encuestadora debe determinar el tamaño n de la muestra aleatoria para escoger de una población de 10,000 electores, ¿qué tan grande debería ser la muestra si se quiere tener una confianza del 95% que el error de estimación no sea superior al 4.8%?. Rp. 400.198≅401.

22. Un auditor toma una muestra aleatoria de 400 cuentas por cobrar y encuentra que 320 de ellas tienen deudas de al menos \$700. Determine el nivel de confianza

a) Si el porcentaje de todas las cuentas por cobrar de al menos \$700 se estima de 75.76% a 84.24%.

b) Si todas las cuentas por cobrar de al menos \$700 de un total de 10,000 cuentas por cobrar se estima en el intervalo [7543, 8457].

Rp. a) 0.966, b) 0.98.

23. Un fabricante estima en 5% la proporción de piezas defectuosos de los 5,000 producidos.

a) Para confirmar tal estimación primero se debe escoger una muestra aleatoria, ¿cuántas piezas debe tener la muestra si se quiere tener una confianza del 95% que el error de la estimación no será superior a 0.047?

b) Se escoge una muestra aleatoria del tamaño calculado en a) , si en ella se encuentran 40 piezas defectuosos, mediante un intervalo de confianza del 95%, ¿se puede inferir que la estimación del fabricante es coherente con la estimación efectuada a partir de la muestra aleatoria?.

Rp. a) $n \cong 400$ b) 0.1 ∓ 0.0282, No, $p=5\%$ no pertenece al I.C..

24. Se quiere estimar p con un error máximo de estimación $e = 0.05$, hallar el tamaño de la muestra necesaria si la población es de tamaño $N=2000$,

Rp. $p=0.5$, $n=333.56 \cong 334$

Diferencia de dos medias y de dos proporciones

25. Para comparar dos métodos de la enseñanza de las matemáticas, se aplicaron a 200 alumnos elegidos al azar el método tradicional y a otra muestra de 250 alumnos el método nuevo resultando las calificaciones promedios respectivos de 13 y 15. Suponga que la varianzas poblacionales respectivas son 9 y 16.

a) Determine un intervalo de confianza del 95% para la diferencia de las medias,

b) ¿Podemos afirmar que el método nuevo es superior al método antiguo? .

Rp. -2 ∓ 0.65, $\mu_1-\mu_2=0 \notin$ IC ademas, $\mu_1-\mu_2<0$, el método nuevo es mejor.

26. Se quiere estimar la diferencia entre los promedios de tiempos (en minutos) que utilizan los hombres y las mujeres para realizar un test de aptitud. Se aplica el test a 20 hombres y 25 mujeres dando las medias respectivas de 110 y 100 puntos. Suponga que las dos poblaciones son normales con varianzas respectivas iguales a 100 y 64.

a) Determine un intervalo de confianza del 98% para la diferencia de las medias,

b) ¿Es válida la afirmación $\mu_1-\mu_2=13$?

Rp. 10 ∓ 5.967, $\mu_1-\mu_2=0 \notin$ IC además, $\mu_1-\mu_2=13 \in$ IC.

27. Se quiere estimar la diferencia entre los promedios de tiempos (en minutos) que utilizan dos operarios para realizar determinada tarea. Suponga que las poblaciones de los dos tiempos se distribuyen normalmente con varianza común. Estime la diferencia entre los dos promedios poblacionales mediante un intervalo de confianza del 95% si el registro de 16 tiempos de cada operario han dado: $\bar{x}_1 = 38$, $\hat{s}_1 = 6$, y $\bar{x}_2 = 35$, $\hat{s}_2 = 4$

Rp. ES=1.8028, gl=30, 3 ∓ 3.68, $\mu_1-\mu_2 \in [-0.68, 6.68]$.

28. Un inversionista hace un estudio para elegir una de dos ciudades del interior del país para abrir un centro comercial. Escoge 21 hogares de la ciudad 1 determinando: $\bar{x}_1 = \$400$, $\hat{s}_1 = \$120$ y escoge 16 hogares de la ciudad 2 calculando: $\bar{x}_2 = \$350$, $\hat{s}_2 = \$60$. Suponga poblaciones normales con varianzas diferentes. Mediante un intervalo de confianza del 95%, ¿se puede afirmar que son iguales los ingresos promedios de las dos ciudades?.

Rp. ES=30.178, gl=31, 20 ∓ 61.08, $\mu_1-\mu_2=0 \in$ IC se acepta, $\mu_1=\mu_2$.

29. Para comparar los gastos promedios mensuales de los alumnos de 2 universidades particulares se escogen dos muestras aleatorias de 10 y 9 alumnos respectivamente resultando los siguientes gastos en dólares:

Muestra 1: 400, 410, 420, 380, 390, 410, 400, 405, 405, 400.
Muestra 2: 390, 395, 380, 390, 400, 380, 370, 390, 380.

Mediante un intervalo de confianza del 95% para la diferencia de los promedios de los gastos mensuales, ¿se puede inferir que los gastos promedios son iguales?. Suponga que ambas poblaciones son normales, independientes, con varianzas desconocidas supuestas iguales .

Rp. Utilizando el *MCEST* se tiene: gl=17, ES=4.726, I.C: 15.89 ∓ 9.97, No.

30. Una agencia de publicidad realizó un estudio para comparar la efectividad de un anuncio en la radio en dos distritos. Después de difundir el aviso, se realizó una encuesta con 900 personas seleccionadas al azar, en cada uno de los distritos, resultando las proporciones 20% y 18% respectivamente. Si de los datos muestrales se infiere que $p_1 - p_2 \in [-0.0162, 0.0562]$, ¿qué nivel de confianza se utilizó?.

Rp. $1-\alpha=0.95$

31. Dos muestras aleatorias de 250 mujeres y 200 hombres indican que 75 mujeres y 80 hombres consumirían un nuevo producto unisex que acaba de salir al mercado. Utilizando un intervalo de confianza del 95%, ¿se puede aceptar que es igual la proporción de preferencias de mujeres y hombres en toda la población?, si no es así, ¿cuál es la relación?

Rp. $-0.1 \mp 1.96(0.0452)$, $[-0.189, -0.011]$, $p_1 \neq p_2$ además es $p_1 < p_2$

Varianzas

32. Se escoge una muestra aleatoria de 13 tiendas y se encuentra que las ventas de la semana de un determinado producto de consumo popular tiene una desviación estándar $\hat{s} = \$6$. Se supone que las ventas del producto tienen una distribución normal. Estimar a) la varianza y b) la desviación estándar poblacional mediante un intervalo de confianza del 95%.

Rp. $(432/23.34) \leq \sigma^2 \leq (432/4.4)$, entonces, $18.51 \leq \sigma^2 \leq 98.18$, y $4.30 \leq \sigma \leq 9.91$

33. Una de las maneras de medir el grado de satisfacción de los empleados de una misma categoría en cuanto a la política salarial, es a través de las desviaciones estándar de sus salarios. La fábrica A afirma ser más homogénea en la política salarial que la fábrica B. Para verificar esa afirmación, se escoge una muestra aleatoria de 10 empleados no especializados de A, y 13 de B, obteniendo las dispersiones $\hat{s}_A = 50$, $\hat{s}_B = 30$ de salario mínimo, ¿cuál sería su conclusión si utiliza un intervalo del 95% para el cociente de varianzas?. Suponga distribuciones normales.

Rp. $0.80 \leq \sigma_1^2 / \sigma_2^2 \leq 10.76$, . Si.

Capítulo 10

PRUEBAS DE HIPOTESIS

Introducción.

El objetivo es dar algunos métodos que se usan para tomar decisiones sobre poblaciones, a partir de los resultados de una muestra aleatoria escogida de esa población. Para llegar a **tomar *decisiones estadísticas* se debe partir de afirmaciones o conjeturas con respecto a la población** en el que estamos interesados. Tales suposiciones, pueden ser verdaderas o no. Una conjetura hecha sobre una población o sobre sus parámetros deberá ser sometida a comprobación experimental con el propósito de saber si los resultados de una muestra aleatoria extraída de esa población, contradicen o no tal conjetura.

10.1 Hipótesis estadísticas.

Definición. Se denomina **hipótesis estadística** a cualquier afirmación o conjetura que se hace acerca de la distribución de una o más poblaciones.

La afirmación o conjetura puede referirse bien a la forma o tipo de distribución de probabilidad de la población o bien referirse al valor o valores de uno o más parámetros de la distribución conocida su forma. En las aplicaciones básicas, se asume dada la forma de la distribución de la población. En este caso, las hipótesis estadísticas consisten en suponer que los parámetros, que definen a la población, toman determinados valores numéricos.

Por **ejemplo**, son hipótesis estadísticas :

1. La longitud media de un tipo de objetos es 10 centímetros.
2. La proporción de objetos defectuosos producidos por cierto proceso nunca es superior al 8%.
3. La varianza de la longitud de cierto tipo de objetos es 0.25 cm^2..
4. Son iguales las medias de dos tipos de mediciones independientes X e Y que se distribuyen normalmente con varianza común σ^2.

Hipótesis simple y compuesta.

Definición. Se denomina **hipótesis simple** a cualquier hipótesis estadística que especifica completamente la distribución de la población, es decir, especifica la forma de la distribución y el valor de su(s) parámetro(s).

Si una hipótesis no especifica completamente la distribución de la población se dice que es una **hipótesis compuesta**.

Por **ejemplo**, la hipótesis que establece que el ingreso mensual promedio de los empleados de cierta ciudad es $\mu = \$500$, suponiendo que los ingresos mensuales se distribuyen según la normal con desviación estándar conocida $\sigma = \$30$, es una hipótesis simple, pues, especifica completamente la distribución de la población. En cambio, si se supone que los ingresos mensuales se distribuyen según la normal con desviación estándar conocida $\sigma = \$30$ y se afirma que el ingreso promedio mensual es $\mu \neq 500$ ó $\mu < 500$ ó $\mu > 500$, entonces, la hipótesis referente a la media es una hipótesis compuesta, pues, no especifica la media de la distribución de la población de los ingresos.

Hipótesis nula y alternativa

Definición. Se denomina **hipótesis nula** y se representa por H_0, a la hipótesis que es aceptada provisionalmente como verdadera y cuya validez será sometida a comprobación experimental. Los resultados experimentales nos permitirán seguir aceptándola como verdadera o si, por el contrario, debemos rechazarla como tal.

Toda hipótesis nula va acompañada de otra hipótesis alternativa.

Definición. Se denomina **hipótesis alternativa** y se representa por H_1 o por H_A, a la hipótesis que se acepta en caso de que la hipótesis nula H_0 sea rechazada. La hipótesis alternativa H_1, es pues una suposición contraria a la hipótesis nula.

Por **ejemplo**, si se asume que θ_0 es un valor del parámetro desconocido θ de una población cuya distribución se supone conocida, entonces son hipótesis nulas y alternativas respectivamente las siguientes afirmaciones:

1) $H_0 : \theta = \theta_0$, y $H_1 : \theta \neq \theta_0$
2) $H_0 : \theta \leq \theta_0$, y $H_1 : \theta > \theta_0$
3) $H_0 : \theta \geq \theta_0$, y $H_1 : \theta < \theta_0$

Prueba de una hipótesis estadística

Para tomar decisiones estadísticas, se requieren de las dos hipótesis: la *hipótesis nula* y la *hipótesis alternativa* referidas a un parámetro θ.

La **prueba de una hipótesis estadística** es un proceso que nos conduce a tomar la decisión de aceptar o rechazar la hipótesis nula H_0, en contraposición de la hipótesis alternativa H_1 y en base a los resultados de una muestra aleatoria seleccionada de la población en estudio.

La hipótesis nula H_0 es la primera hipótesis que se plantea, y debe ser establecida de manera que *especifique un valor* θ_0 del parámetro θ en estudio. Por esta razón, algunos autores plantean la hipótesis nula $H_0 : \theta = \theta_0$ aun para los casos 2) y 3) del ejemplo anterior .

La aceptación de una hipótesis significa que los datos de la muestra no proporcionan evidencia suficiente para refutarla. El rechazo significa que los datos de la muestra lo refutan.

Tipos de pruebas de hipótesis.

El **tipo de prueba** depende básicamente de la hipótesis alternativa H_1.

Se denomina **prueba de una cola** a toda prueba de hipótesis donde la alternativa H_1 es unilateral. Si la alternativa es bilateral, la prueba se denomina **prueba de dos colas.**

La prueba de hipótesis $H_0 : \theta = \theta_0$ contra $H_1 : \theta \neq \theta_0$ se denomina **prueba bilateral o de dos colas.**

La prueba de hipótesis $H_0 : \theta = \theta_0$ contra $H_1 : \theta > \theta_0$ se denomina **prueba unilateral de cola a la derecha**.

La prueba de hipótesis $H_0 : \theta = \theta_0$ contra $H_1 : \theta < \theta_0$ se denomina **prueba unilateral de cola a la izquierda**

Errores tipo I y tipo II, y Nivel de significación

Al tomar la decisión de *aceptar* o *rechazar* la hipótesis nula $H_0 : \theta = \theta_0$ en base a los resultados obtenidos de una muestra aleatoria seleccionada de la población en estudio; hay cuatro posibles situaciones que determinan si la decisión tomada es correcta o incorrecta, como se muestra en la tabla 10-1.

Definición. Se denomina **error tipo I,** al error que se comete al rechazar una hipótesis nula H_0 cuando ésta realmente es verdadera.

Definición. Se denomina **error tipo II,** al error que se comete al aceptar una hipótesis nula H_0 cuando en realidad es falsa.

La probabilidad de cometer un error tipo I se denota por α. Entonces,

$$\alpha = P[\text{error tipo I}] = P[\text{rechazar } H_0 \text{ cuando } H_0 \text{ es verdadera}].$$

La probabilidad de cometer un error tipo II se denota por β. Entonces,

$$\beta = P[\text{error tipo II}] = P[\text{aceptar } H_0 \text{ cuando } H_0 \text{ es falsa}].$$

Tabla 10-1

Decisión	H_0 verdadera	H_0 falsa
Rechazar H_0	Error tipo I Probab: α	Decisión correcta Probab: $1-\beta$
Aceptar H_0	Decisión correcta Probab: $1-\alpha$	Error tipo II Probab: β

Definición. Se denomina **nivel de significación** de una prueba de hipótesis a la probabilidad de cometer un error de tipo I.

Definición. La **potencia** de una prueba es la probabilidad de tomar la decisión acertada de, rechazar H_0 cuando ésta es falsa o de aceptar H_1 cuando ésta es verdadera La potencia de una prueba es calculada por $1-\beta$.

El nivel de significación se fija previamente por lo general en $\alpha = 0.05$ o $\alpha = 0.01$. Si para un valor dado de α, se rechaza la hipótesis H_0, entonces se dice que los resultados muestrales obtenidos, no sólo son diferentes por efectos del azar, si no que son realmente **significativamente diferentes** al nivel $\alpha \times 100\%$, es decir; se espera que de 100 resultados muestrales en $\alpha \times 100\%$ de las veces se rechazará la hipótesis nula H_0 cuando realmente es verdadera.

Para una muestra aleatoria de tamaño n seleccionada de la población en estudio, si α aumenta, entonces β disminuye, y si β aumenta, entonces α disminuye. Por supuesto, en todo proceso de toma de decisiones sobre hipótesis estadísticas, es deseable disminuir las probabilidades de cometer esos dos tipos de errores. Este problema es tratado ampliamente en otros textos de estadística.

Región crítica y regla de decisión

Después de plantear la hipótesis nula H_0 y su correspondiente alternativa H_1 referentes a un parámetro θ, y especificado el tamaño α del nivel de significación de la prueba de H_0 contra H_1, se deberá **determinar una estadística** Θ correspondiente al parámetro, cuya distribución muestral se conozca. Por ejemplo, si las hipótesis H_0 y H_1 se expresan en términos de la media poblacional μ,

entonces se seleccionará la media muestral $\overline{X}$ como la estadística apropiada para efectuar la prueba.

Si se supone que la hipótesis nula $H_0: \theta = \theta_0$ es verdadera; entonces, la distribución de probabilidad de la estadística Θ queda bien definida por esta hipótesis, ya que esta hipótesis especifica completamente la distribución.

En la distribución de probabilidad fijada por la hipótesis H_0: $\theta = \theta_0$ se establece la ***regla de decisión*** de acuerdo con la cual se rechazará o por el contrario se aceptará la hipótesis H_0. El rechazo de la hipótesis nula H_0 implica la aceptación de H_1.

La regla de decisión implica la división de la distribución muestral del estadístico Θ de la prueba en dos partes mutuamente excluyentes: la ***región de rechazo*** o ***región crítica*** ***(R.C.)*** de H_0, y la ***región de aceptación*** *(R.A.)* o *no rechazo* de H_0. Esta división depende de la hipótesis alternativa H_1, del nivel de significación α y de la distribución muestral del estadístico.

Por **ejemplo,** supongamos que se tiene una población normal $N(\mu, 9)$ con **varianza conocida** $\sigma^2 = 9$ y que se trata de **probar o *docimar*** la hipótesis nula $H_0: \mu = 70$ contra $H_1: \mu > 70$.

Dado que $\overline{X}$ es un buen estimador de μ utilizaremos esta estadística para determinar la región crítica y la regla de decisión de esta prueba. **Puesto que estamos interesados en la discriminación entre** $\mu = 70$ **y valores de** $\mu > 70$ **parece razonable que debamos rechazar** H_0 **si** $\overline{X} - 70$ **es muy grande, esto es si** $\overline{X} > K$ **, siendo *K* un *valor crítico* que vamos a determinar.**

Si **se supone verdadera** la hipótesis $H_0: \mu = 70$, entonces, la distribución de la media $\overline{X}$ es normal con media $\mu = 70$ y desviación estándar $\sigma = 3$.

En consecuencia la distribución de

$$Z = \frac{\overline{X} - 70}{3/\sqrt{n}}$$

es normal $N(0,1)$.

Para una muestra aleatoria de tamaño $n = 40$ y la probabilidad de error tipo I, $\alpha = 0.5$ se tiene (ver figura 10.1).

$$0.05 = P[\text{rechazar } H_0 \text{ cuando } H_0 \text{ es veradera}] = P[\overline{X} > K/\mu = 70]$$

$$0.05 = P\left[\frac{\overline{X} - 70}{3/\sqrt{40}} > \frac{K - 70}{3/\sqrt{40}}\right] = P\left[Z > \frac{K - 70}{0.474}\right].$$

De la tabla normal $N(0,1)$ se obtiene:

$$\frac{K-70}{0.474}=1.645\text{, luego } K=70+1.645\times 0.474=70.78\text{.}$$

Por tanto, la *región crítica* en el rango de variación de $\overline{X}$ es el intervalo

$$R.C.=]70.78,+\infty[\text{.}$$

La ***regla de decisión*** es: si $\overline{x}$ es el valor de $\overline{X}$ obtenido a partir de una muestra aleatoria de tamaño 40, se rechazará H_0 si $\overline{x}>70.78$.

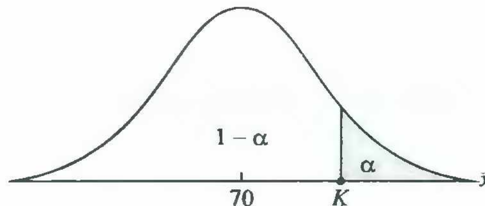


Figura 10.1: Región crítica cola a la derecha en la variable $\overline{X}$

Procedimiento de la prueba de hipótesis

Previamente debe formularse el problema estadístico, determinar la variable en estudio y el método estadístico adecuado para la solución del problema. El procedimiento general de la prueba de una hipótesis de parámetro θ se resume en los siguientes pasos:

1) Formular la *hipótesis* nula $H_0:\theta=\theta_0$ y la hipótesis alternativa adecuada $H_1:\theta\neq\theta_0$ ó $H_1:\theta>\theta_0$ ó $H_1:\theta<\theta_0$
2) Especificar el tamaño α del *nivel de significación*.
3) Seleccionar la *estadística* apropiada a usar en la prueba.
4) Establecer la *regla de decisión*, determinando la región crítica de la prueba.
5) *Calcular* el valor del estadístico de la prueba a partir de los datos de la muestra.
6) Tomar la *decisión* de rechazar la hipótesis H_0 si el valor del estadístico de la prueba está en la región crítica. En caso contrario, no rechazar H_0.

10.2 Pruebas de hipótesis acerca de la media μ: Varianza σ^2 supuesta conocida

Sea $\overline{X}$ la media de una muestra aleatoria de tamaño n seleccionada de una población con media μ y varianza σ^2 supuestamente conocida.

Si la *población es normal* $N(\mu, \sigma^2)$, entonces, la distribución de la estadística $\overline{X}$ es exactamente normal $N(\mu, \sigma^2/n)$ para cualquier valor de n $(n \geq 2)$. Si la *población no es normal*, pero el tamaño de la muestra es suficientemente grande ($n \geq 30$), entonces, la distribución de $\overline{X}$ es aproximadamente normal $N(\mu, \sigma^2/n)$. Entonces,

La estadística para la prueba acerca de μ con varianza σ^2 conocida es

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

cuya distribución es exacta o aproximadamente normal estándar $N(0,1)$, según sea la población normal o no

Si se supone verdadera la hipótesis nula: $H_0 : \mu = \mu_0$, la estadística especificada por esta hipótesis es entonces:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$, dado el nivel de significación α, en la distribución de $Z = (\overline{X} - \mu_0)/(\sigma / \sqrt{n})$, que es normal $N(0,1)$, se determina el valor $z_{1-\alpha/2}$ tal que la probabilidad de rechazar H_0 cuando se supone verdadera sea (figura 10.2)

$$P[Z < -z_{1-\alpha/2}] = \alpha/2 \quad \text{o} \quad P[Z > z_{1-\alpha/2}] = \alpha/2.$$

En consecuencia, ***la región crítica en el rango de variación de* Z** es:

$$R.C. = \{Z < -z_{1-\alpha/2} \ \text{ o } \ Z > z_{1-\alpha/2}\}.$$

Por otro lado, la probabilidad de aceptar H_0 cuando se supone verdadera es:

$$P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1 - \alpha.$$

Resultando la ***región de aceptación***: $R.A. = \{-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}\}$.

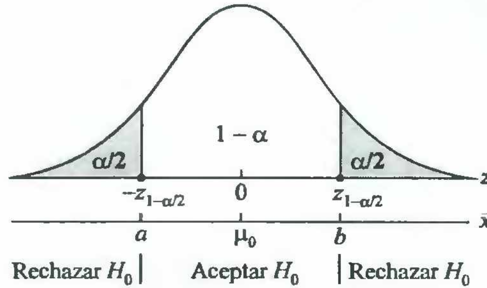


Figura 10.2: Región crítica bilateral en escalas z y $\overline{x}$

Regla de decisión es: Si $z_k = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ es un valor de Z obtenido de la muestra, entonces, se rechazará H_0 con riesgo igual a α, si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A.$).
No se rechazará H_0 en caso contrario (figura 10.2).

Si se rechaza H_0 se dice que el *valor z_k es significativo* con un riesgo cuyo valor es α.

NOTA. (Región crítica en $\overline{X}$)

Si se sustituye $Z = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ en RC resulta ***la región crítica en el rango de variación de*** $\overline{X}$:

$$R.C. = \{\overline{X} < a \quad \text{o} \quad \overline{X} > b\}.$$

donde $a = \mu_0 - z_{1-\alpha/2}(\sigma/\sqrt{n})$, y $b = \mu_0 + z_{1-\alpha/2}(\sigma/\sqrt{n})$

La región de aceptación es el intervalo en $\overline{X}$:

$$R.A. = [a \leq \overline{X} \leq b].$$

La *regla de decisión* es: Si $\bar{x}$ es el valor de $\overline{X}$ obtenido a partir de una muestra aleatoria, se rechazará H_0 con un riesgo α, si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A.$).
No se rechazará H_0 en caso contrario (figura 10.2).

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu > \mu_0$, dado el nivel de significación α, en la distribución de $Z = (\overline{X} - \mu_0)/(\sigma / \sqrt{n})$ que es normal $N(0,1)$, se determina el valor $z_{1-\alpha}$ tal que (figura 10.3),

$$P[Z > z_{1-\alpha} \,/\, H : \mu = \mu_0 \text{ verdadera}] = \alpha$$

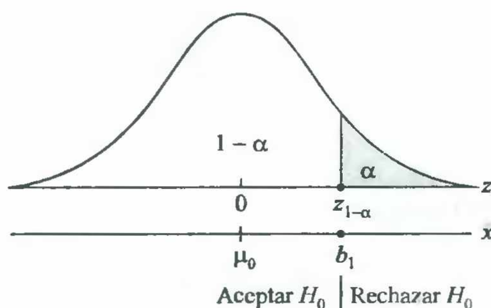


Figura 10.3: Región crítica cola a la derecha en escalas z y $\overline{x}$

En consecuencia, *la región crítica en el rango de variación de* Z es:

$$R.C. = \{Z > z_{1-\alpha}\}$$

La *región de aceptación* es: $R.A. = \{Z \leq z_{1-\alpha}\}$.

La *regla de decisión* es: Si $z_k = (\overline{x} - \mu_0)/(\sigma / \sqrt{n})$ es un valor de Z obtenido a partir de una muestra, se rechazará H_0 si $z_k \in R.C.$ (o si $z_k \notin R.A.$).
No se rechazará H_0 en caso contrario (figura 10.3).

NOTA. (Región crítica en $\overline{X}$)

Si se sustituye $Z = (\overline{X} - \mu_0)/(\sigma / \sqrt{n})$ en RC resulta la región crítica en el rango de variación de $\overline{X}$:

$$R.C. = \{\overline{X} > b_1\}$$

donde, $b_1 = \mu_0 + z_{1-\alpha}(\sigma / \sqrt{n})$

La *región* de aceptación es el intervalo:

$$R.A. = \{\overline{X} \leq b_1\}$$

La *regla de decisión* es: Siendo $\bar{x}$ el valor de $\bar{X}$ obtenido a partir de una muestra aleatoria de tamaño *n*, se rechazará H_0 con un riesgo α, si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A.$). No se rechazará H_0 en caso contrario.(figura 10.3).

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu < \mu_0$, dado el nivel de significación α, en la distribución de $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$, se puede determinar el valor $z_{1-\alpha}$ tal que (figura 10.4)

$$P[Z < -z_{1-\alpha} / H : \mu = \mu_0 \text{ verdadera}] = \alpha$$

En consecuencia, l*a región crítica en el rango de variación de* ***Z*** es:

$$R.C. = \{Z < -z_{1-\alpha}\}$$

Región de aceptación es: $R.A. = \{Z \geq -z_{1-\alpha}\}$.

Regla de decisión : Si $z_k = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ es un valor de Z obtenido a partir de una muestra, se rechazará H_0 con un riesgo α, si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A.$).
No se rechazará H_0 en caso contrario (figura 10.4).

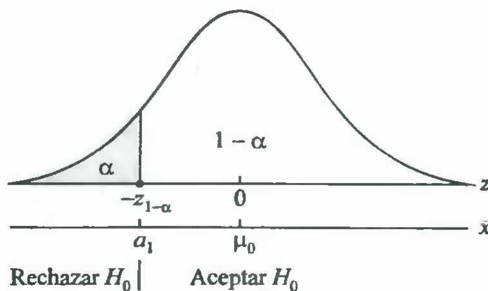


Figura 10.4: Región crítica cola a la izquierda en escalas z y $\bar{x}$

NOTA. (Región crítica en $\bar{X}$)

Si se sustituye $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ en RC se obtiene la región crítica en el rango de variación de $\bar{X}$:

$$RC = \{\bar{X} < a_1\}$$

donde, $$a_1 = \mu_0 - z_{1-\alpha}(\sigma/\sqrt{n})$$

Región de aceptación: $RA = \{\overline{X} \geq a_1\}$.

Regla de decisión es: Si $\bar{x}$ es un valor de $\overline{X}$ obtenido a partir de una muestra aleatoria de tamaño *n*, se rechazará H_0 con un riesgo α si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A.$).

No se rechazará H_0 en caso contrario (figura 10.4).

EJEMPLO 10.1.

Un determinado proceso de empaquetar un producto está controlado, si el peso medio del producto empaquetado es 400 gramos. Si en una muestra aleatoria de 100 paquetes del producto se ha encontrado que el peso medio es de 395 gramos, ¿Se podría concluir que el proceso está fuera de control al nivel de significación 5%?. Suponga que el peso de los productos empaquetados se distribuye normalmente con desviación estándar de 20 gramos.

SOLUCION.

Sea *X* la variable aleatoria definida como el peso de los paquetes del producto. Se supone que la distribución de *X* es $N(\mu, (20)^2)$.

1. *Hipótesis*: $H_0 : \mu = 400$ (el proceso está controlado).

   $H_1 : \mu \neq 400$ (el proceso está fuera de control).

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística*: Población normal con varianza conocida, la estadística es

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

   cuya distribución es normal $N(0,1)$.

4. *Región crítica*: Si se supone verdadera la hipótesis nula H_0, para $\alpha = 0.5$ y la alternativa bilateral, en la distribución de $Z = (\overline{X} - 400)/(20/\sqrt{100})$, se encuentra el valor crítico:

$$z_{1-\alpha/2} = z_{0.975} = 1.96$$

Luego, la región crítica en la variable *Z* está dada por:

$$RC = \{Z < -1.96 \quad \text{o} \quad Z > 1.96\}$$

5. *Cálculos*: De los datos se tiene:

$$n = 100,\ \bar{x} = 395,\ \sigma = 20.$$

$$z_k = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{395 - 400}{2} = -2.5.$$

6. *Decisión*: Puesto que $z_k = -2.5 \in R.C$, debemos rechazar H_0 y concluir con un riesgo de 5%, que el proceso de empaquetar no está controlado.

NOTA. En el rango de variación de $\bar{X}$, la región crítica es:

$$R.C. = \{\bar{X} < 400 - 1.96 \times 2 \quad \text{o} \quad \bar{X} > 400 + 1.96 \times 2\}$$

$$R.C. = \{\bar{X} < 396.08 \quad \text{o} \quad \bar{X} > 403.92\}$$

El hecho que $\bar{x} = 395 \in R.C.$, se debe rechazar H_0 y concluir con un riesgo de 5%, que el proceso de empaquetar no está controlado.

NOTA. (Regla de decisión en Intervalo de confianza)

La **prueba bilateral** de la hipótesis nula $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$ a un nivel de significación dado α, **equivale a calcular el intervalo de confianza** (*I.C.*) de $(1-\alpha)\times 100\%$ para el parámetro μ y luego rechazar la hipótesis nula $H_0 : \mu = \mu_0$ si es que $\mu_0 \notin I.C.$

En efecto, si $\bar{x}$ es un valor de $\bar{X}$, no se rechazará $H_0 : \mu = \mu_0$ si el valor

$$z_k \in R.A. = [-z_{1-\alpha/2},\ z_{1-\alpha/2}], \text{ donde } z_k = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$$

o, si

$$-z_{1-\alpha/2} \leq \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}$$

Esto es, no se rechazará $H_0 : \mu = \mu_0$ si

$$\bar{x} \in R.A. = [\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \ \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

o equivalentemente si μ_0 se encuentra dentro del intervalo de confianza (*I.C.*) del $(1-\alpha)\times 100\%$ para μ:

$$\mu_0 \in I.C. = [\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Por tanto, se rechazará H_0 con riesgo α si,

$$\bar{x} \notin R.A. \quad \text{o si} \quad \mu_0 \notin I.C$$

Por **ejemplo**, en el ejemplo 10.1, para $\alpha = 0.05$ se tiene:

$$I.C. = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = [391.08, 398.92]$$

$$R.A. = \left[\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = [396.08, 403.92].$$

Dado que $\mu_0 = 400 \notin I.C.$ (o que $\bar{x} = 395 \notin R.A.$) se debe rechazar H_0 con un riesgo del 5%.

NOTA. (Método del valor *P* en la prueba)

Otra forma de establecer la regla de decisión, en estadística aplicada, es calculando el valor *P*, a partir del *valor absoluto* de $z_k = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$, (que se obtiene de la muestra), de manera que

a) $P = P[Z < -z_k] + P[Z > z_k] = 2P[Z > z_k]$ (para dos colas).

b) $P = P[Z > z_k]$ (cola a la derecha).

c) $P = P[Z < -z_k]$ (cola a la izquierda).

Si el valor de $P < \alpha$, entonces, se rechazará H_0. No se rechazará H_0, en caso contrario.

Las pruebas de hipótesis con el paquete estadístico *MCEST* contienen el método del valor *P*.

En el **ejemplo 10.1**, el valor absoluto de z_k es igual a 2.5, entonces,

$$P = 2P[Z > z_k] = 2P[Z > 2.5] = 2(0.0062) = 0.0124.$$

Dado que $P = 0.0124 < \alpha = 0.05$, se debe rechazar H_0, con un riesgo $\alpha = 0.05$. Observar que z_k es significativo en un valor mucho menor que $\alpha = 0.05$ y que este valor de z_k sólo ocurrirá en 124 casos de 10,000 experimentos.

Una región crítica de tamaño 0.0124 es muy pequeña y, por lo tanto, es poco probable que se cometa error tipo I.

EJEMPLO 10.2.

Al estudiar si conviene tener o no una sucursal en la ciudad de Tarapoto, la gerencia de una gran tienda comercial de Lima, establece el siguiente criterio para tomar una decisión: Abrir la sucursal sólo si el ingreso promedio familiar mensual en dicha ciudad es no menos de $500 y no abrirla en caso contrario. Si una muestra aleatoria de 100 ingresos familiares de esa ciudad ha dado una media de $480.

a) ¿Cuál es la decisión a tomar al nivel de significación del 5%?.
b) ¿Con qué probabilidad la prueba anterior detecta la diferencia igual a $30 en el promedio de ingresos y por debajo de lo que se indica en la hipótesis nula?.
c) Calcular la potencia de la prueba si el ingreso promedio realmente es $464.

Suponga que la distribución de los ingresos tiene una desviación estándar igual a $80.

SOLUCION.

Sea X la variable aleatoria que representa los ingresos familiares mensuales de los pobladores de Tarapoto.

a) 1. *Hipótesis*: $H_0 : \mu = 500$ (ó $H_0 : \mu \geq 500$) (se abre la sucursal).

$H_1 : \mu < 500$ (no se abre la sucursal).

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística*: Población no normal, $n = 100$, $\sigma = 80$, por teorema central del límite la estadística apropiada es:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

cuya distribución es aproximadamente normal $N(0,1)$.

4. *Región crítica*: Si se supone verdadera la hipótesis nula H_0 para $\alpha = 0.5$ y la alternativa unilateral cola a la izquierda, en la distribución de $Z = (\overline{X} - 500)/(80/\sqrt{100})$, se encuentra el valor crítico

$$z_\alpha = z_{0.05} = -1.645 .$$

Luego, la región crítica en la variable Z es: $RC = \{Z < -1.645\}$

5. *Cálculos*: De la muestra se tiene, $\overline{x} = 480$

$$z_k = \frac{\overline{x} - 500}{80/\sqrt{100}} = \frac{480 - 500}{8} = -2.5,$$

6. *Decisión*: Dado que $z_k = -2.5 \in R.C$, debemos rechazar H_0 y concluir con no abrir la sucursal en la ciudad de Tarapoto.

b) La región crítica de H_0 es:

$$RC = \{Z < -1.645\} = \{\overline{X} < 500 - 1.645 \times 8\} = \{\overline{X} < 486.84\}$$

La prueba anterior detecta la diferencia igual a \$30 en el promedio de ingresos por debajo de lo que se indica en la hipótesis nula si se rechaza H_0 cuando $\mu = 470$. Entonces,

$$P[\overline{X} < 486.84 / \mu = 470] = P\left[Z < \frac{486.84 - 470}{80/\sqrt{100}}\right] = P[Z < 2.11] = 0.9826$$

c) La potencia de la prueba es: $1 - \beta = 1 - 0.0021 = 0.9979$, donde la probabilidad β de aceptar H_0 cuando realmente es $\mu = 464$ (error tipo II) es:

$$\beta = P[\overline{X} \geq 486.84 / \mu = 464] = P\left[Z \geq \frac{486.84 - 464}{80/\sqrt{100}}\right] = P[Z \geq 2.86]$$

$$\beta = P[Z \geq 2.86] = 0.0021.$$

EJEMPLO 10.3. (Tamaño de la muestra)

Suponga que X es una población normal con media μ (desconocida) y con varianza σ^2 conocida. Dadas las probabilidades α y β de cometer errores tipo I y tipo II respectivamente, determinar el tamaño n de la muestra requerida para probar las hipótesis simples

$$H_0 : \mu = \mu_0 \text{ contra } H_1 : \mu = \mu_1 \text{ donde } \mu_1 < \mu_0 .$$

SOLUCION.

Sea K el punto crítico en la variable $\overline{X}$, de la prueba unilateral cola a la izquierda de H_0 contra H_1 (figura 10.5), entonces,

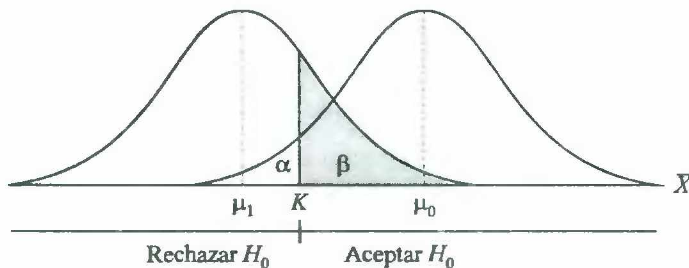


Figura 10.5: Tamaño de muestra en prueba de hipotesis

$$\alpha = P[\text{error tipo I}] = P[\text{rechazar } H_0 \ / \ H_0 : \mu = \mu_0 \text{ es verdadera}]$$

$$\alpha = P\left[\overline{X} < K / \mu = \mu_0\right] = P\left[\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{K - \mu_0}{\sigma/\sqrt{n}}\right] = P\left[Z < \frac{K - \mu_0}{\sigma/\sqrt{n}}\right]$$

de donde resulta (por ser α menos del 50% del área total y $z_\alpha = -z_{1-\alpha}$)

$$\frac{K - \mu_0}{\sigma/\sqrt{n}} = -z_{1-\alpha}, \qquad K = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \tag{1}$$

De igual manera,

$$\beta = P[\text{error tipo II}] = P[\text{aceptar } H_0 \ / \ H_0 : \mu = \mu_0 \text{ es falsa}]$$

$$\beta = P[\text{aceptar } H_0 \ / \ H_1 : \mu = \mu_1 \text{ es verdadera}]$$

$$\beta = P[\overline{X} \geq K / \mu = \mu_1] = P\left[\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} \geq \frac{K - \mu_1}{\sigma/\sqrt{n}}\right] = P\left[Z \geq \frac{K - \mu_1}{\sigma/\sqrt{n}}\right]$$

de donde resulta,

$$\frac{K - \mu_1}{\sigma/\sqrt{n}} = z_{1-\beta}, \qquad K = \mu_1 + z_{1-\beta}\frac{\sigma}{\sqrt{n}} \tag{2}$$

Resolviendo para n las ecuaciones (1) y (2). resulta

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(\mu_0 - \mu_1)^2}.$$

NOTA.

Si la prueba unilateral es de cola a la derecha, esto es si se prueba:

$$H_0 : \mu = \mu_0 \text{ contra } H_1 : \mu = \mu_1, \text{ donde } \mu_0 < \mu_1$$

el tamaño n de la muestra requerida es también:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(\mu_0 - \mu_1)^2}.$$

Si la prueba es bilateral, esto es, si se prueba: $H_0 : \mu = \mu_0$ contra $H_1 : \mu = \mu_1$, $\mu_0 \neq \mu_1$, el tamaño n de la muestra requerida es

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2}{\left(\mu_0 - \mu_1\right)^2}.$$

EJEMPLO 10.4.

El gerente de ventas de una compañía afirma que sus vendedores venden en promedio diariamente $1500. Diseñe una prueba para esta hipótesis si se quiere que el riesgo sea 1.5% de rechazar la afirmación cuando realmente es verdadera y el riesgo sea de 0.75% de aceptar la afirmación cuando realmente es $1523 el promedio de las ventas diarias. Suponga que la distribución de las ventas es normal con una desviación estándar de $50

SOLUCION.

Sean las hipótesis: $H_0 : \mu = 1500$ y $H_1 : \mu = 1523$.

Si se supone verdadera la hipótesis nula $H_0 : \mu = 1500$, entonces la distribución de $Z = \dfrac{\overline{X} - 1500}{50/\sqrt{n}}$ es normal $N(0,1)$.

Luego, para $\alpha = 0.015 = P[\text{Error tipo I}]$, en esta distribución se tiene:

$$0.015 = P[\text{rechazar } H_0 \ / \ H_0 : \mu = 1500 \text{ es verdadera}]$$

$$0.015 = P\left[\overline{X} > K/\mu = 1500\right] = P\left[Z > \frac{K - 1500}{50/\sqrt{n}}\right]$$

de donde resulta: $\dfrac{K - 1500}{50/\sqrt{n}} = +2.17, \qquad K = 1500 + \dfrac{2.17 \times 50}{\sqrt{n}}$

También, si se supone verdadera la hipótesis alternativa $H_1 : \mu = 1523$, entonces, la distribución de $Z = \dfrac{\overline{X} - 1500}{50/\sqrt{n}}$ es normal $N(0,1)$.

Luego, para $\beta = 0.0075 = P[\text{Error tipo II}]$, en esta distribución se tiene:

$$0.0075 = P[\text{aceptar } H_0 \ / \ H_1 : \mu = 1523\,]$$

$$0.0075 = P\left[\overline{X} \le K/\mu = 1523\right] = P\left[Z \le \frac{K - 1523}{50/\sqrt{n}}\right]$$

$$\frac{K - 1523}{50/\sqrt{n}} = -2.43, \qquad K = 1523 - \frac{2.43 \times 50}{\sqrt{n}}.$$

Luego de: $K = 1500 + \dfrac{2.17 \times 50}{\sqrt{n}}$ y de $K = 1523 - \dfrac{2.43 \times 50}{\sqrt{n}}$, se obtiene:

$$\sqrt{n} = 10, \qquad n = 100$$

Con el valor de $n = 100$, se obtiene el crítico de la prueba: $K = 1510.85$.
Si $\bar{x}$ es un valor de media de la muestra de $n = 100$ casos, se rechazará la hipótesis nula H_0 si $\bar{x} > 1510.85$. En caso contrario, no se debe rechazar H_0.

10.3 Pruebas de hipótesis acerca de la media μ: Varianza σ^2 supuesta desconocida

A) Población no normal

Si la población no tiene distribución normal y si la varianza es desconocida, para probar hipótesis acerca de la media μ, sólo si, el tamaño de la muestra es grande ($n \geq 30$), **se suele utilizar la estadística**:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

cuya distribución es aproximadamente $N(0,1)$. La desviación estándar σ se estima puntualmente por $\hat{s}$.

Luego, las regiones críticas de la pruebas de $H_0 : \mu = \mu_0$ contra cualquiera de las tres alternativas $H_1 : \mu > \mu_0$ ó $H_1 : \mu < \mu_0$ ó $H_1 : \mu \neq \mu_0$ son las mismas (aproximadamente) de la sección 10.2.

B) Población normal

Si la población tiene distribución normal $N(\mu, \sigma^2)$, donde μ y σ^2 son parámetros desconocidas, para $n \geq 2$ la estadística de la prueba acerca de la media μ es:

$$T = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}$$

cuya distribución es ***t*-Student** con $n - 1$ grados de libertad.

Si se supone verdadera la hipótesis nula: $H_0 : \mu = \mu_0$, la estadística especificada por esta hipótesis es entonces, ahora:

$$T = \frac{\overline{X} - \mu_0}{\hat{S}/\sqrt{n}}$$

NOTA. La estructura de la prueba es idéntica que en el caso de σ conocida, salvo que el valor de σ se estima por $\hat{s}$ y la distribución normal estándar se sustituye por la distribución *t* de Student con $n-1$ grados de libertad.

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$, dado el nivel de significación α, en la distribución de $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n}) \sim t(n-1)$, se determinan los valores $\pm t_{1-\alpha/2, n-1}$, tales que la probabilidad de rechazar H_0 cuando se supone verdadera sea (figura 10.6)

$$P[T < -t_{1-\alpha/2, n-1}] = \alpha/2 \quad \text{o} \quad P[T > t_{1-\alpha/2, n-1}] = \alpha/2 .$$

Luego, la ***región crítica en el rango de variación de*** *T* es:

$$R.C. = \{T < -t_{1-\alpha/2, n-1} \text{ o } T > t_{1-\alpha/2, n-1}\}$$

La región de aceptación es el intervalo

$$R.A. = \{-t_{1-\alpha/2, n-1} \leq T \leq t_{1-\alpha/2, n-1}\}$$

Regla de decisión: Se rechazará H_0 con riesgo α, si $t_k \in R.C.$ (o, si $t_k \notin R.A$). No se rechazará H_0 en caso contrario.

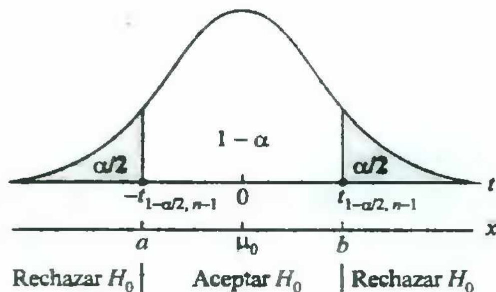


Figura 10.6: Región crítica bilateral en escalas *t* y $\overline{x}$

NOTA. Si se sustituye $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n})$ en *R.C* se obtiene:

la ***región crítica en el rango de variación*** *de* $\overline{X}$:

$$R.C. = \{\overline{X} < a \quad \text{o} \quad \overline{X} > b\}$$

donde $a = \mu_0 - t_{1-\alpha/2,n-1}(\hat{s}/\sqrt{n})$, y $b = \mu_0 + t_{1-\alpha/2,n-1}(\hat{s}/\sqrt{n})$

Regla de decisión: Siendo $\bar{x}$ el valor de $\overline{X}$ obtenido a partir de una muestra aleatoria de tamaño *n*, se rechazará H_0 con un riesgo α. si $\bar{x} \in R.C.$ (o si $\bar{x} \notin R.A. = (R.C.)^c$). No se rechazará H_0 en caso contrario (figura 10.6).

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu > \mu_0$, dado el nivel de significación α, en la distribución de $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n}) \sim t(n-1)$, se determina el valor $t_{1-\alpha,n-1}$ tal que: (figura 10.7)

$$P[T > t_{1-\alpha,n-1} / H_0 : \mu = \mu_0 \text{ verdadera}] = \alpha$$

Luego, *la región crítica en el rango de variación de T* es:

$$RC = \{T > t_{1-\alpha,n-1}\}$$

La región de aceptación es el intervalo:

$$RA = \{T \leq t_{1-\alpha,n-1}\}$$

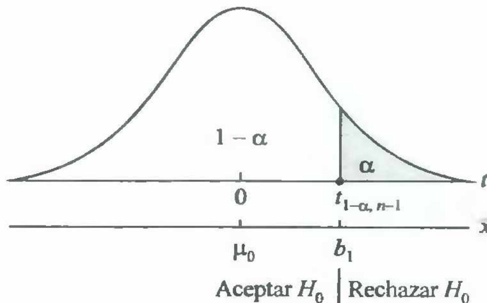


Figura 10.7: Región crítica cola a la derecha en escalas *t* y $\overline{X}$

Regla de decisión: Se rechazará H_0 si $t_k \in R.C.$ (o si $t_k \notin R.A.$). No se rechazará H_0 en caso contrario.

NOTA. La región crítica en $\overline{X}$ (figura 10.7) es: $RC = \{\overline{X} > b_1\}$, donde $b_1 = \overline{X} < \mu_0 + t_{1-\alpha, n-1}(\hat{s}/\sqrt{n})$

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu = \mu_0$ contra $H_1 : \mu < \mu_0$, dado el nivel de significación α, en la distribución de $T = (\overline{X} - \mu_0)/(\hat{s}/\sqrt{n}) \sim t(n-1)$ se determina el valor $t_{1-\alpha, n-1}$, tal que; (figura 10.8)

$$P[T < -t_{1-\alpha, n-1} / H_0 : \mu = \mu_0 \text{ verdadera}] = \alpha$$

Luego, *la región crítica en el rango de variación de* T es:

$$RC = \{T < -t_{1-\alpha, n-1}\}$$

La región de aceptación es el intervalo:

$$RA = \{T \geq -t_{1-\alpha, n-1}\}$$

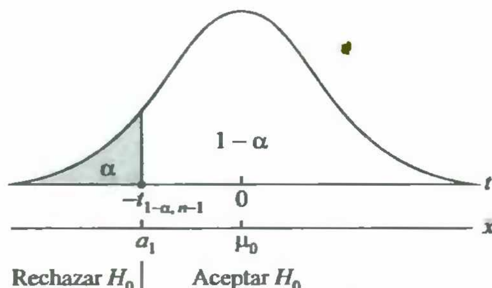


Figura 10.8: Región crítica cola a la izquierda en escalas t y $\overline{X}$

Regla de decisión: Se rechazará H_0 si $t_k \in R.C.$ (o si $t_k \notin R.A.$). No se rechazará H_0 en caso contrario (figura 10.8).

NOTA. La región crítica en $\overline{X}$ (figura 10.8) es $RC = \{\overline{X} < a_1\}$, donde $a_1 = \overline{X} < \mu_0 - t_{1-\alpha, n-1}(\hat{s}/\sqrt{n})$

EJEMPLO 10.5.

Las cajas de cierto tipo de cereal procesadas por una fábrica deben tener un contenido promedio de 160 gramos. Por una queja ante el defensor del consumidor de que tales cajas de cereal tienen menos contenido, un inspector tomó una muestra aleatoria de 10 cajas encontrando los siguientes pesos de cereal en gramos:

157, 157, 163, 158, 161, 159, 162, 159, 158, 156

¿Es razonable que el inspector multe al fabricante?. Utilice un nivel de significación del 5% y suponga que los contenidos tienen distribución normal.

SOLUCION.

Sea X la variable aleatoria que representa los pesos de las cajas del cereal. Se supone que la distribución X es normal con media μ y varianza σ^2 desconocidas.

1) *Hipótesis*: $H_0: \mu = 160$ (No multa al fabricante)

$H_1: \mu < 160$ (Multa al fabricante)

2. *Nivel de significación* $\alpha = 0.05$
3. *Estadística*: Población normal, con varianza desconocida y $n = 10$.
   Si $H_0: \mu = 160$ es verdadera, la estadística es

$$T = \frac{\overline{X} - 160}{\hat{s}/\sqrt{n}}$$

que se distribuye según una t-Student con 9 grados de libertad.

4. *Región crítica*: Con el nivel de significación $\alpha = 0.05$ y para una prueba de hipótesis unilateral cola a la izquierda, en la tabla de probabilidades de t-Student se encuentra: $t_{0.95,9} = 1.833$.

   Consecuentemente, la región crítica es: $RC = \{T < -1.833\}$
5. *Cálculos*: De los datos de la muestra se obtiene:

$$n = 10,\ \bar{x} = 159,\ \hat{s} = 2.309, \text{ error estándar: } \hat{s}/\sqrt{n} = 0.73$$

$$t_k = \frac{\bar{x} - 160}{\hat{s}/\sqrt{n}} = \frac{159 - 160}{0.73} = -1.37.$$

6. *Decisión*: Dado que $t_k = -1.37 \notin R.C$, debemos aceptar H_0 y concluir que la media de los ingresos quincenales no ha variado.

NOTA. Utilizando el paquete de computo *MCEST*, se encuentra la probabilidad $P = P[T < -1.37] = 0.1012$, por lo que debemos aceptar H_0.

EJEMPLO 10.6.

Un fabricante produce un cable de alambre de cierto tipo, que tiene una resistencia a la ruptura no mayor de 300 kg.. Se descubre un proceso nuevo y más barato que desea emplearse, siempre que el cable así producido tenga una resistencia media a la ruptura mayor de 300 kg.. Si una muestra aleatoria de 36 cables producidos con el nuevo proceso ha dado una media 304.5 kg. y una desviación estándar $\hat{s} = 15$ kg. ¿Debería el fabricante adoptar el nuevo proceso, si está dispuesto a asumir un error tipo I del 5%?.

Suponga que la distribución de la resistencia a la ruptura es:

a) Normal, b) Desconocida no normal.

SOLUCION.

Sea X la resistencia a la ruptura.

Se debe probar la hipótesis nula $H_0 : \mu \leq 300$ contra $H_1 : \mu > 300$ al nivel de significación del 5% y a partir de una muestra de tamaño: $n = 36$.

a) Si X tiene distribución normal, la varianza σ^2 es desconocida y si $H_0 : \mu = 300$ es realmente cierta, la estadística de la prueba es:

$$T = \frac{\overline{X} - 300}{\hat{s}/\sqrt{n}} \sim t(35)$$

Para $\alpha = 0.05$ y una prueba unilateral de cola a la derecha, en la distribución $t(35)$ se encuentra el valor crítico: $t_{0.950,35} = 1.69$.

Luego, la región crítica es: $RC = \{T > 1.69\}$

De los datos de la muestra, se tiene:

$$n = 36,\ \bar{x} = 304.5,\ \hat{s} = 15,\ \text{error estándar: } \hat{s}/\sqrt{n} = 2.5$$

$$t_k = \frac{\bar{x} - 300}{\hat{s}/\sqrt{n}} = \frac{304.5 - 300}{2.5} = 1.8.$$

Dado que, $t_k = 1.8 \in R.C.$, se debe rechazar H_0.

b) Si X no se distribuye en forma normal, la varianza σ^2 es desconocida, $n \geq 30$, y si $H_0 : \mu = 300$ es realmente cierta, la estadística es:

$$Z = \frac{\overline{X} - 300}{\hat{s}/\sqrt{n}}$$

que se distribuye aproximadamente normal $N(0,1)$.

Para el nivel de significación $\alpha = 0.05$ y para una prueba unilateral de cola a la derecha, en la distribución de Z se encuentra: $z_{1-\alpha} = z_{0.95} = 1.645$. Luego, la región crítica es:

$$RC = \{Z > 1{,}645\}.$$

De la muestra aleatoria resulta:

$$z_k = \frac{\bar{x} - 300}{\hat{s}/\sqrt{n}} = \frac{304.5 - 300}{2.5} = 1.8.$$

Ya que $z_k = 1.8 \in R.C.$, debemos rechazar H_0 y concluir que conviene adoptar el nuevo proceso.

10.4 Pruebas de hipótesis acerca de una varianza

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n, seleccionada de una población normal con media μ y varianza σ^2, parámetros desconocidos, y sea la varianza muestral,

$$\hat{S}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

Entonces, la variable aleatoria,

$$X = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

tiene distribución chi-cuadrado con $n-1$ grados de libertad. Esta estadística se utiliza para probar hipótesis acerca de una varianza.

Si se supone verdadera la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$, la estadística es:

$$X = \frac{(n-1)\hat{S}^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Su valor $x_k = \dfrac{(n-1)\hat{s}^2}{\sigma_0^2}$ que resulta de la muestra aleatoria, se utiliza para la prueba de H_0, contra una alternativa unilateral o bilateral.

1) Prueba bilateral o de dos colas

Si se prueba $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 \neq \sigma_0^2$, dado un nivel de significación α, en la distribución $\chi^2(n-1)$ se determinan los valores $\chi^2_{\alpha/2,n-1}$ y $\chi^2_{1-\alpha/2,n-1}$ (figura 10.9) tales que la probabilidad de rechazar la hipótesis nula H_0 cuando realmente es verdadera es igual a:

$$P[X < \chi^2_{\alpha/2,n-1}] = \alpha/2 \quad \text{o} \quad P[X > \chi^2_{1-\alpha/2,n-1}] = \alpha/2.$$

La *Región crítica* de la prueba, es entonces,

$$R.C. = \{X < \chi^2_{\alpha/2,n-1} \quad \text{o} \quad X > \chi^2_{1-\alpha/2,n-1}\}.$$

La *Regla de decisión* es rechazar H_0 con un riesgo α, si $x_k \in R.C.$ (o si $x_k \notin R.A. = [\chi^2_{\alpha/2,n-1}, \chi^2_{1-\alpha/2,n-1}]$). No rechazar H_0 en caso contrario.

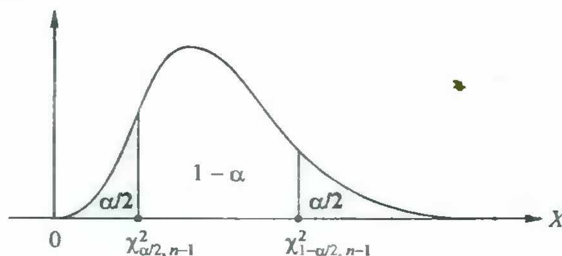


Figura 10.9: Región crítica para la prueba de H_0: $\sigma^2 = \sigma_0^2$ contra H_1: $\sigma^2 \neq \sigma_0^2$

2) Contraste unilateral de cola a la derecha..

Si se prueba $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 > \sigma_0^2$, dado un nivel de significación α, en la distribución $\chi^2(n-1)$ se determina el valor $\chi^2_{1-\alpha,n-1}$ (figura 10.10) tal que la probabilidad de rechazar la hipótesis nula H_0 cuando realmente es verdadera es igual a:

$$P[X > \chi^2_{1-\alpha,n-1}] = \alpha.$$

Luego, la *región crítica* es: $R.C. = \{X > \chi^2_{1-\alpha,n-1}\}$.

La *regla de decisión* es: rechazar H_0 al nivel α si $x_k \in R.C.$ (o si $x_k \notin R.A. = \{X \leq \chi^2_{1-\alpha,n-1}\}$). No rechazar H_0, en caso contrario.

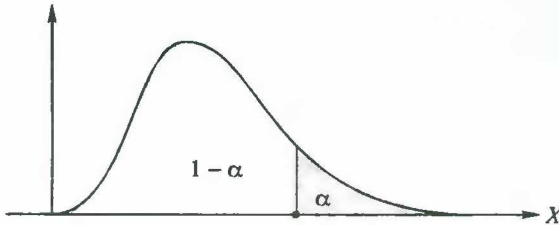


Figura 10.10: Región crítica para la prueba de H_0: $\sigma^2 = \sigma_0^2$ contra H_1: $\sigma^2 > \sigma_0^2$

3) Contraste unilateral cola a la izquierda.

Si la prueba es de $H_0: \sigma^2 = \sigma_0^2$ contra $H_1: \sigma^2 < \sigma_0^2$, dado un nivel de significación α, en la distribución $\chi^2(n-1)$ se determina el valor $\chi^2_{\alpha,n-1}$ (figura 10.11) tal que la probabilidad de rechazar la hipótesis nula H_0 cuando realmente es verdadera es igual a:

$$P[X < \chi^2_{\alpha,n-1}] = \alpha.$$

Luego, la *región crítica* es: $R.C. = \{X < \chi^2_{\alpha,n-1}\}$

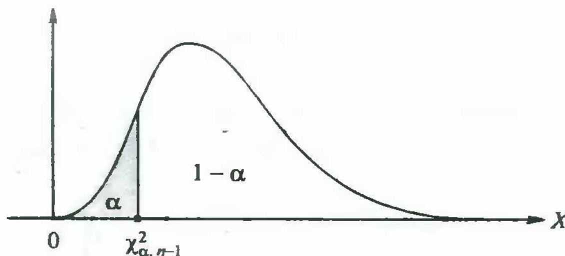


Figura 10.11 :Región crítica para la prueba de H_0: $\sigma^2 = \sigma_0^2$ contra H_1: $\sigma^2 < \sigma_0^2$

Regla de decisión: rechazar H_0 si $x_k \in R.C.$ (o si $x_k \notin R.A. = \{X \geq \chi^2_{\alpha,n-1}\}$). No rechazar H_0 en caso contrario.

EJEMPLO 10.7.

En un proceso de fabricación, se plantea la hipótesis que la desviación estándar de las longitudes de cierto tipo de tornillo es 2.0 mm En una muestra de diez tornillos elegidos al azar del proceso de producción se han encontrado las siguientes longitudes en milímetros:

$$71, 66, 64, 72, 69, 67, 70, 68, 65, 69.$$

Con estos datos, ¿se justifica la suposición que la desviación estándar verdadera es 2.00 mm?

Use el nivel de significación $\alpha = 0.05$, y suponga que la distribución de las longitudes es normal.

SOLUCION.

1. *Hipótesis*: $H_0 : \sigma^2 = 4$ contra $H_1 : \sigma^2 \neq 4$
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Población normal, con $n = 10$, y suponiendo verdadera la hipótesis $H_0 : \sigma^2 = 4$, la estadística

$$X = \frac{(n-1)\hat{S}^2}{4}$$

   se distribuye como chi-cuadrado con 9 grados de libertad.
4. *Región crítica*: Para $\alpha = 0.05$ y para un contraste bilateral, en la tabla chi-cuadrado se encuentran los siguientes valores críticos:

$$\chi^2_{\alpha/2, n-1} = \chi^2_{0.025, 9} = 2.70$$

$$\chi^2_{1-\alpha/2, n-1} = \chi^2_{0.975, 9} = 19.02.$$

   Luego, la región crítica es: $R.C. = \{X < 2.70 \text{ o } X > 19.02\}$.
5. *Cálculos*: De los datos de la muestra resulta $\hat{s}^2 = 6.77$, entonces,

$$x_k = \frac{(n-1)\hat{s}^2}{4} = \frac{9\hat{s}^2}{4} = \frac{9(6.77)}{4} = 15.23.$$

6. *Decisión*: Como $x_k = 15.23 \notin R.C.$ no se debe rechazar H_0 y concluimos que la varianza de la población es igual a 4 gr^2.

10.5 Pruebas de hipótesis acerca de la razón de dos varianzas

Sean $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de dos muestras aleatorias *independientes* de tamaños respectivos n_1 y n_2, escogidas de **dos poblaciones normales** con **varianzas** respectivas σ_1^2 y σ_2^2. Entonces, la estadística,

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2}$$

tiene distribución de probabilidad F con grados de libertad $r_1 = n_1 - 1$ y $r_2 = n_2 - 1$. Esto es, $F \sim F(r_1, r_2)$. Esta estadística se utiliza para probar igualdad de varianzas.

Si se supone verdadera la hipótesis nula $H_0: \sigma_1^2 = \sigma_2^2$ o $\sigma_1^2/\sigma_2^2 = 1$ la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} \sim F(r_1, r_2)$$

Su valor $f_k = \frac{\hat{s}_1^2}{\hat{s}_2^2}$ que resulta de dos muestras aleatorias, se utiliza para probar la hipótesis nula H_0 contra cualquier alternativa unilateral o bilateral

Observar que para obtener la estadística *F*, no se requiere asumir que las dos poblaciones tengan igual promedio.

1) **Prueba bilateral.** Si se prueba $H_0: \sigma_1^2 = \sigma_2^2$ contra $H_1: \sigma_1^2 \neq \sigma_2^2$, dado un nivel de significación α, en la distribución de $F(r_1, r_2)$ se encuentran los valores $f_{\alpha/2, r_1, r_2}$ y $f_{1-\alpha/2, r_1, r_2}$ tales que la probabilidad de rechazar H_0 cuando realmente es verdadera es igual a (figura 10.12).

$$P[F < f_{\alpha/2, r_1, r_2}] = \alpha/2 \quad \text{o} \quad P[F > f_{1-\alpha/2, r_1, r_2}] = \alpha/2.$$

Luego la *región crítica* de la prueba es

$$R.C. = \{F < f_{\alpha/2, r_1, r_2} \text{ o } F > f_{1-\alpha/2, r_1, r_2}\}.$$

La *regla de decisión* es rechazar H_0 si $f_k \in R.C.$ o si $f_k \notin R.A. = [f_{\alpha/2, r_1, r_2}, f_{1-\alpha/2, r_1, r_2}]$. No rechazar H_0 en caso contrario.

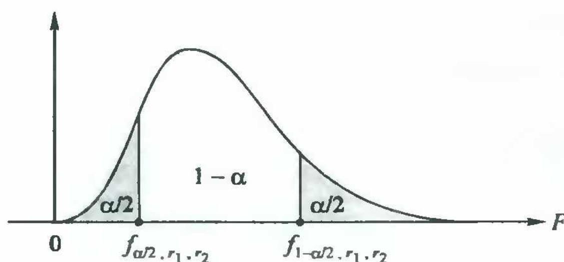


Figura 10.12: Región crítica de la prueba de H_0: $\sigma_1^2 = \sigma_2^2$ contra H_1: $\sigma_1^2 \neq \sigma_2^2$

2) **Prueba Unilateral cola derecha.** Si se prueba $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 > \sigma_2^2$, dado el nivel de significación α, en la distribución $F(r_1, r_2)$ se determina el valor $f_{1-\alpha, r_1, r_2}$ tal que la probabilidad de rechazar H_0 cuando realmente es verdadera es igual a;

$$P[F > f_{1-\alpha, r_1, r_2}] = \alpha.$$

Luego, la región crítica de la prueba es: $R.C. = \{F > f_{1-\alpha, r_1, r_2}\}$

La *regla de decisión* es: Rechazar H_0 si $f_k \in R.C.$ o si $f_k \notin R.A. =]-\infty, f_{1-\alpha, r_1, r_2}]$. No rechazar H_0 en caso contrario

3) **Prueba Unilateral cola izquierda.** Si se prueba $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 < \sigma_2^2$, dado el nivel de significación α, en la distribución $F(r_1, r_2)$ se determina el valor f_{α, r_1, r_2} tal que la probabilidad de rechazar H_0 cuando realmente es verdadera es igual a;

$$P[F < f_{\alpha, r_1, r_2}] = \alpha$$

Luego, la región crítica es: $R.C. = \{F < f_{\alpha, r_1, r_2}\}$.

La *regla de decisión* es: Rechazar H_0, si $f_k \in R.C.$ o si $f_k \notin R.A. = \{F \geq f_{\alpha, r_1, r_2}\}$. No rechazar H_0 en caso contrario.

EJEMPLO 10.8.

Una compañía diseña un nuevo proceso de moldeo para reducir la variabilidad en el diámetro de las piezas producidas. Se cree que la varianza del nuevo proceso es menor que la varianza del proceso antiguo. Para una muestra de 8 piezas del proceso antiguo y una muestra de 6 piezas del proceso nuevo se obtienen los siguientes diámetros en milímetros:

Antiguo: 17, 23, 21, 18, 22, 20, 21, 19.
Nuevo: 13, 16, 14, 12, 15, 14.

¿Confirman estos datos que la varianza de los diámetros con el nuevo proceso es menor que con el proceso antiguo?.
Suponga poblaciones normales y use $\alpha = 0.05$.

SOLUCION.

Sean X_1 y X_2 las variables que representan los diámetros de las piezas con el proceso antiguo y nuevo respectivamente. Las dos poblaciones se distribuyen normalmente con varianzas desconocidas respectivas σ_1^2 y σ_2^2.

1. *Hipótesis*: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 > \sigma_2^2$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Poblaciones normales. Suponiendo verdadera la hipótesis nula H_0, para $n_1 = 8$ y $n_2 = 6$, la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

que se distribuye como $F(7,5)$.

4. *Región crítica*. Para $\alpha = 0.05$ y la prueba unilateral cola a la derecha, en la distribución de $F(7,5)$, la región crítica es

$$R.C. = \{ F > 4.88 \}.$$

5. *Cálculos*. De los datos de la muestra se obtiene:

$$\hat{s}_1^2 = 4.125, \quad \hat{s}_2^2 = 2 \quad \text{y} \quad f_k = \hat{s}_1^2 / \hat{s}_2^2 = 4.125/2 = 2.0625.$$

6. *Decisión*. Como $f_k = 2.065 \notin R.C.$ no se debe rechazar H_0 y concluir que las dos varianzas son iguales.

NOTA Con el paquete de computo *MCEST* se obtiene la probabilidad: $P[F>2.0625]=0.221$,con la que se toma la decisión que las dos varianza son iguales.

10.6 Pruebas de hipótesis acerca de dos medias

10.6.1 Pruebas de hipótesis acerca de dos medias: Varianzas σ_1^2 y σ_2^2 supuestas conocidas

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones independientes, con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 respectivas supuestas conocidas.

Si las *dos poblaciones son normales*, entonces, las estadísticas $\overline{X}_1$ y $\overline{X}_2$ tienen respectivamente distribución normal $N(\mu_1, \sigma_1^2/n_1)$ y $N(\mu_2, \sigma_2^2/n_2)$ para $n_1 \geq 2$, y $n_2 \geq 2$. Luego, la estadística $\overline{X}_1 - \overline{X}_2$ tiene distribución exactamente normal $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Si las dos *poblaciones no* son *normales* pero n_1 y n_2 son suficientemente grandes ($n_1 \geq 30$ y $n_2 \geq 30$), entonces, $\overline{X}_1 - \overline{X}_2$ tiene distribución aproximadamente normal $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Luego, según sean las dos poblaciones normales o no, la estadística

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

tiene distribución exactamente o aproximadamente normal $N(0,1)$.

Si suponemos *verdadera* la hipótesis nula $H_0: \mu_1 = \mu_2$ o $\mu_1 - \mu_2 = 0$, la estadística es:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

Su valor $z_k = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ que resulta de dos muestras, se utiliza para probar H_0 contra cualquiera de las hipótesis alternativas $H_1: \mu_1 \neq \mu_2$ ó $H_1: \mu_1 > \mu_2$ ó $H_1: \mu_1 < \mu_2$.

La estructura de la prueba es similar a los casos descritos en la sección 10.2, usando la distribución de Z.

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$, la región crítica **en el rango de variación** de Z es:

$$R.C. = \{Z < -z_{1-\alpha/2} \text{ o } Z >_{1-\alpha/2}\}.$$

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$, la región crítica en la variación de Z es,

$$R.C. = \{Z > z_{1-\alpha}\}.$$

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 < \mu_2$, la región crítica en los valores de Z es:

$$R.C. = \{Z < - z_{1-\alpha}\}.$$

NOTA. Cuando las hipótesis son de la forma

1) $H_0 : \mu_1 - \mu_2 = d_0$ contra $H_1 : \mu_1 - \mu_2 \neq d_0$
2) $H_0 : \mu_1 - \mu_2 = d_0$ contra $H_1 : \mu_1 - \mu_2 > d_0$
3) $H_0 : \mu_1 - \mu_2 = d_0$ contra $H_1 : \mu_1 - \mu_2 < d_0$

La estadística de la prueba es,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - d_0}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

cuya distribución es exactamente o aproximadamente normal $N(0,1)$, según sean las dos poblaciones normales o no.

EJEMPLO 10.9.

Un fabricante quiere comparar dos marcas de máquinas, A y B; para fabricar un tipo de artículo. Observa dos muestras aleatorias de 60 artículos procesados por A y B respectivamente y encuentra que las medias respectivas son 1,230 y 1,190 segundos.

Suponga $\sigma_1 = 120$ y $\sigma_2 = 90$ segundos.

a) Al nivel de significación del 5%, ¿se puede inferir que la máquina B es más rápida que la máquina A?.

b) Al nivel de significación del 5%, ¿se puede inferir que la media de B es menor que la media de A en menos de 7 segundos?.

c) ¿En cuánto deberían incrementarse los tamaños de las muestras de cada proceso para que la diferencia observada de 40 segundos en los tiempos promedios muestrales de A menos B sea significativa al nivel α=1%?

SOLUCION.

Sean X_1 y X_2 los tiempos de proceso con las máquinas A y B respectivamente y μ_1 y μ_2 sus medias respectivas.

Se desconocen las distribuciones de probabilidades de X_1 y X_2, pero las muestras son grandes.

a) 1. *Hipótesis*: $H_0: \mu_1 = \mu_2$ contra $H_1: \mu_1 > \mu_2$

2. *Nivel de significación*: $\alpha = 0.05$.

3. *Estadística*: Si se supone verdadera la hipótesis H_0 y para muestras grandes, la estadística apropiada es:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

cuya distribución es aproximadamente normal estándar $N(0,1)$.

4. *Región crítica*: Para $\alpha = 0.05$ y una prueba unilateral de cola a la derecha, en la distribución de Z se encuentra el valor $z_{0.9500} = 1.645$. Luego, la región crítica es,

$$R.C. = \{Z > 1.645\}.$$

5. *Cálculos*: De los datos se tiene

$$n_1 = n_2 = 60, \quad \bar{x}_1 = 1{,}230, \ \bar{x}_2 = 1190, \quad \sigma_1 = 120 \text{ y } \sigma_2 = 90$$

$$E.S\text{=Error estándar } = \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)} = 19.365$$

$$z_k = \frac{\bar{x}_1 - \bar{x}_2}{ES} = \frac{1{,}230 - 1{,}190}{19.365} = 2.07.$$

6. *Decisión*: Ya que $z_k = 2.07 \in R.C.$, debemos rechazar H_0 y concluir que el equipo B utiliza menor tiempo en el proceso de fabricación,

b) Se debe probar $H_0 : \mu_1 - \mu_2 = 7$ contra $H_1 : \mu_1 - \mu_2 > 7$.
Si H_0 es verdadera, la estadística de la prueba es

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - 7}{\sqrt{\left(\sigma_1^2/n_1\right) + \left(\sigma_2^2/n_2\right)}} \sim N(0,1)$$

La región crítica de la prueba unilateral cola derecha al nivel $\alpha = 0.05$ es: la misma del caso a)

$$R.C. = \{Z > 1.645\}.$$

$$z_k = \frac{(\bar{x}_1 - \bar{x}_2) - 7}{ES} = \frac{(1230 - 1190) - 7}{19.365} = 1.7.$$

Ya que $z_k = 1.7 \in R.C.$, debemos rechazar H_0 y concluir que la máquina B utiliza un tiempo promedio menos de 7 segundos debajo del tiempo promedio de A.

c) Sea n el tamaño de cada una de las dos muestras tomadas de los artículos procesados por las máquinas A y B.
Si la hipótesis nula H_0 es verdadera, y si además

$$\bar{x}_1 - \bar{x}_2 = 1230 - 1190 = 40,$$

se tiene
$$z_k = \frac{40}{\sqrt{\dfrac{120^2 + 90^2}{n}}} = 0.27\sqrt{n},$$

Para $\alpha = 0.01$ y una prueba unilateral de cola a la derecha, en la distribución de Z se halla:

$$z_{1-\alpha} = z_{0.99} = 2.33.$$

La región crítica es: $R.C = \{Z > 2.33\}$
La diferencia observada de 40 en las medias de las muestras será significativa al nivel de 1%, si

$$0.27\sqrt{n} \in R.C.$$

Esto es, si $0.27\sqrt{n} > 2.33$, $\sqrt{n} > 8.63$, $n \geq 75$.

De aquí que se debe incrementar cada muestra en al menos: $75 - 60 = 15$ casos

10.6.2 Pruebas de hipótesis acerca de dos medias: Varianzas σ_1^2 y σ_2^2 supuestas desconocidas

A) Poblaciones no normales

Si las dos muestras aleatorias independientes de tamaños n_1 y n_2 se seleccionan respectivamente de dos poblaciones cuyas *distribuciones son no normales* con varianzas σ_1^2 y σ_2^2 supuestas desconocidas , entonces, siempre que los tamaños de las muestras sean grandes; $n_1 \geq 30$ y $n_2 \geq 30$ los parámetros σ_1 y σ_2 se estiman respectivamente por $\hat{s}_1$ y $\hat{s}_2$.

Para probar la hipótesis nula $H_0 : \mu_1 - \mu_2 = 0$ contra una alternativa bilateral o unilateral, se utiliza la estadística:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}}$$

que se distribuye aproximadamente normal $N(0,1)$.

Las regiones críticas y las reglas de decisión para las pruebas de la hipótesis nula $H_0 : \mu_1 - \mu_2 = 0$ (ó $H_0 : \mu_1 - \mu_2 = d_0$) contra una alternativa unilateral o bilateral **son las mismas del método con varianzas conocidas.**

B) Poblaciones normales

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias y $\hat{S}_1^2$ y $\hat{S}_2^2$ las varianzas de dos **muestras aleatorias independientes** de tamaños n_1 y n_2 respectivamente seleccionadas de *dos poblaciones normales* con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2 desconocidas.

B1) Varianzas desconocidas supuesta iguales: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Si las poblaciones son normales, independientes, y con varianzas desconocidas supuestas iguales, entonces, la estadística

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_c^2}{n_1} + \frac{\hat{S}_c^2}{n_2}}}$$

tiene distribución *t*-student con $n_1 + n_2 - 2$ grados de libertad, en donde la varianza común:

$$\hat{S}_c^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

es un *estimador insesgado de la varianza común* σ^2.

Si la hipótesis nula $H_0 : \mu_1 = \mu_2$ es verdadera, entonces, la estadística

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\hat{S}_c^2}{n_1} + \dfrac{\hat{S}_c^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

Su valor

$$t_k = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

que resulta de dos muestras aleatorias, se utiliza para probar H_0 contra una alternativa unilateral o bilateral.

La estructura de la prueba es similar a los casos descritos en la sección 10.3, usando la distribución de *t*.

1) Prueba bilateral o de dos colas

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$, la región crítica es el intervalo:,

$$R.C. = \{T < -t_{1-\alpha/2,\, n_1+n_2-2} \text{ o } T > t_{1-\alpha/2,\, n_1+n_2-2}\}$$

2) Prueba unilateral de cola a la derecha

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 > \mu_2$, la región crítica es el intervalo:

$$R.C. = \{T > t_{1-\alpha,\, n_1+n_2-2}\}$$

3) Prueba unilateral de cola a la izquierda

Si se prueba $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 < \mu_2$, la región crítica es el intervalo:

$$R.C. = \{T < -t_{1-\alpha,\, n_1+n_2-2}\}$$

B2) Varianzas desconocidas supuestas distintas $\sigma_1^2 \neq \sigma_2^2$

Si las varianzas de las dos poblaciones normales independientes son desconocidas supuestas diferentes, entonces, la estadística

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2}},$$

tiene distribución t-student con r grados de libertad, siendo

$$r = \frac{\left[\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2\right]^2}{\dfrac{\left[\hat{s}_1^2/n_1\right]^2}{n_1 - 1} + \dfrac{\left[\hat{s}_2^2/n_2\right]^2}{n_2 - 1}}$$

Dado que r rara vez es un entero, se redondea al entero más cercano.

Si la hipótesis nula $H_0: \mu_1 = \mu_2$ se supone verdadera, entonces

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2}} \sim t(r)$$

Su valor $t_k = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2}}$ que resulta de dos muestras aleatorias independientes, se utiliza para probar H_0 contra una alternativa unilateral o bilateral.

Las regiones críticas y las reglas de decisión son similares a los del caso B1, pero con r grados de libertad.

EJEMPLO 10.10.

Una medicina A es aplicada a 10 pacientes aquejados de cierta enfermedad. Otra medicina B es aplicada a otros 9 pacientes aquejados de la misma enfermedad. Los tiempos de recuperación de los pacientes, en días, fueron los siguientes:

Medicina A: 6, 5, 6, 7, 4, 7, 6, 4, 3, 6.

Medicina B: 7, 6, 7, 9, 5, 8, 7, 6, 8.

Utilizando un nivel de significación del 5% y suponiendo poblaciones normales.

a) ¿Se podría concluir que las varianzas poblacionales son iguales?
b) ¿Se puede aceptar la hipótesis nula que son iguales las medias de los tiempos de tratamiento de las dos medicinas?.
c) ¿Cuál de las medicinas es más eficaz?

SOLUCION.

Sean X_1 y X_2 las variables aleatorias que representan los tiempos en días de tratamiento de las medicinas A y B respectivamente. Se supone que $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$.

a) Prueba de la homogeneidad de varianzas

1. *Hipótesis*: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 \neq \sigma_2^2$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Poblaciones normales. Suponiendo verdadera la hipótesis nula H_0, para $n_1 = 10$ y $n_2 = 9$, la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

que se distribuye como $F(9, 8)$.

4. *Región crítica*. Para $\alpha = 0.05$ y una prueba bilateral, en la distribución de $F(9,8)$ se encuentran:

$$f_{0.975,9,8} = 4.36 \quad \text{y} \quad f_{0.025,9,8} = 1/f_{0.975,8,9} = 1/4.10 = 0.244.$$

Luego, la región crítica está dada por:

$$R.C. = \{F < 0.244 \text{ o } F > 4.36\}.$$

5. *Cálculos*. De los datos de la muestra se obtiene:

$$\hat{s}_1^2 = 1.822, \quad \hat{s}_2^2 = 1.5 \quad \text{y} \quad f_k = \hat{s}_1^2 / \hat{s}_2^2 = 1.215.$$

6. *Decisión*. Como $f_k = 1.215 \notin R.C.$ Se debería aceptar H_0 y concluir que las varianzas de los tiempos de A y B son iguales.

b) Prueba de la diferencia de las dos medias.

1. *Hipótesis*: $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística de la prueba*: Si se supone H_0 verdadera y dado que hay prueba de que las varianzas poblacionales son iguales, la estadística apropiada es:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\hat{S}_c^2}{n_1} + \frac{\hat{S}_c^2}{n_2}}}$$

que se distribuye según una t–Student con $n_1 + n_2 - 2 = 17$ grados de libertad.

4. *Región crítica*: Para $\alpha = 0.05$ y una prueba de hipótesis bilateral, en la distribución $t(17)$ se encuentra $t_{1-\alpha/2, n_1+n_2-2} = t_{0.975,17} = 2.110$.

   La región crítica en la variación de T es

$$R.C. = \{ T < -2.110 \text{ o } T > 2.110 \}.$$

5. *Cálculos*: De los datos se tiene:

$$n_1 = 10,\ \bar{x}_1 = 5.4,\ \hat{s}_1^2 = 1.822, \quad n_2 = 9,\ \bar{x}_2 = 7.0,\ \hat{s}_2^2 = 1.5.$$

$$\hat{s}_c = \frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2} = \frac{9(1.822) + 8(1.5)}{10 + 9 - 2} = 1.67$$

$$\text{Error estándar} = \sqrt{\frac{\hat{s}_c^2}{n_1} + \frac{\hat{s}_c^2}{n_2}} = 0.594$$

$$t_k = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_c^2}{n_1} + \frac{\hat{s}_c^2}{n_2}}} = \frac{5.4 - 7.0}{0.594} = -2.694.$$

6. *Decisión*: Ya que $t_k = -2.694 \in R.C.$, debemos rechazar H_0 y concluir que los promedios de los tiempos de tratamientos con las medicinas A y B son diferentes.

c) Como las medias de las dos poblaciones son diferentes, planteamos las hipótesis:

$H_0 : \mu_1 = \mu_2$ (Ambas medicinas son iguales).

$H_1 : \mu_1 < \mu_2$ (Medicina A es mejor que B).

Con $\alpha = 0.05$ y 17 grados de libertad, para la prueba unilateral de cola a la izquierda se encuentra el valor crítico: $t_{0.95,17} = 1.740$. Luego, la región crítica es:

$$R.C. = \{ T < -1.740 \}.$$

Como $t_k = -2.694 \in R.C.$, debemos rechazar H_0 y concluir que la medicina A más eficaz que la medicina B.

NOTA. Con el paquete *MCEST* para la prueba de varianzas se obtiene: $P[F>1.215]=0.397$. Dado que 0.397>0.05 se infiere que las varianzas poblacionales son iguales al nivel 5%.

También para la prueba de dos medias se obtiene $P[T > 2.694]=0.007$. Luego, 0.007 es la significación para una prueba unilateral y 2(0.007)=0.014 para una prueba bilateral. Por lo que se debe rechazar H_0 en una prueba unilateral o bilateral al 5%.

EJEMPLO 10.11.

El encargado de compras de una compañía tiene que escoger entre dos marcas de máquinas A y B, para procesar cierto producto. Por cuestiones de precio el encargado desearía comprar la marca A a no ser que haya evidencias de que la máquina B es más veloz. Se le permitió operar los dos tipos de máquinas durante un periodo de prueba, escogiendo al azar luego, los tiempos en segundos de 10 objetos procesados por cada máquina:

Máquina A: 55, 56, 57, 56, 58, 53, 54, 59, 60, 57
Máquina B: 50, 51, 42, 50, 40, 60, 53, 44, 48, 58

Utilizando un nivel de significación del 5% y suponiendo poblaciones de tiempos normales.

a) ¿Se podría concluir que las varianzas poblacionales son iguales?
b) ¿Qué tipo de máquina debería comprar la empresa?.

SOLUCION.

Sean X_1 y X_2 las variables aleatorias que representan los tiempos empleados por las máquinas A y B respectivamente. Se sabe que $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$.

a) Prueba de la homogeneidad de varianzas

1. *Hipótesis*: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 \neq \sigma_2^2$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Poblaciones normales. Suponiendo verdadera la hipótesis nula H_0, para $n_1 = 10$ y $n_2 = 10$, la estadística de la prueba es:

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

que se distribuye como $F(9, 9)$.

4. *Región crítica*. Para $\alpha = 0.05$ y una prueba bilateral, en la distribución de $F(9,9)$ se encuentra:

$$f_{0.975,9,9} = 4.03 \quad \text{y} \quad f_{0.025,9,9} = 1/f_{0.975,9,9} = 1/4.03 = 0.248.$$

Luego, la región crítica está dada por:

$$R.C. = \{F < 0.248 \text{ o } F > 4.03\}.$$

5. *Cálculos*. De los datos de la muestra se obtiene:

$$\hat{s}_1^2 = 4.722, \quad \hat{s}_2^2 = 41.822 \quad \text{y} \quad f_k = \hat{s}_1^2/\hat{s}_2^2 = 4.722/41.822 = 0.1129.$$

6. *Decisión*. Como $f_k = 0.1129 \in R.C.$ se debe rechazar H_0 y concluir que las varianzas de A y B son diferentes.

b) Prueba de diferencia de las dos medias:

1. *Hipótesis*: $H_0 : \mu_1 = \mu_2$ (Los promedios de tiempos de A y B son iguales)
   $H_0 : \mu_1 > \mu_2$ (La máquina B tiene mejor tiempo promedio)
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística de la prueba*: Si se supone H_0 verdadera y dado que hay prueba de que las varianzas poblacionales son diferentes, la estadística apropiada es:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\hat{S}_1^2/n_1 + \hat{S}_2^2/n_2}},$$

que se distribuye según una t-Student con r grados de libertad donde:

$$r = \frac{\left[\hat{s}_1^2/n_1 + \hat{s}_2^2/n_2\right]^2}{\frac{\left[\hat{s}_1^2/n_1\right]^2}{n_1 - 1} + \frac{\left[\hat{s}_2^2/n_2\right]^2}{n_2 - 1}} = \frac{[4.72/10 + 41.82/10]^2}{\frac{[4.72/10]^2}{9} + \frac{[41.82/10]^2}{9}} = 11.007 \cong 11.$$

4. *Región crítica*: Para $\alpha = 0.05$ y una prueba unilateral de cola a la derecha, en la distribución $t(11)$, se encuentra $t_{1-\alpha,\, r} = t_{0.95,11} = 1.796$. La región crítica en la variación de T es

$$R.C. = \{T > 1.796\}$$

5. *Cálculos*: De los datos se tiene también:

$$\bar{x}_1 = 56.5, \quad \bar{x}_2 = 49.6, \quad \bar{x}_1 - \bar{x}_2 = 6.9,$$

$$\text{E.S} = \text{Error estándar} = \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} = \sqrt{\frac{4.72}{10} + \frac{41.82}{10}} = 2.1574$$

$$t_k = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} = \frac{6.9}{2.1574} = 3.198.$$

6. *Decisión*: Ya que $t_k = 3.198 \in R.C.$, debemos rechazar H_0. Concluimos que se debe adquirir la máquina B.

NOTA. Con el paquete *MCEST* se obtiene: $P[F > 8.856]=0.003$ con la que **se rechaza** la afirmación de que las varianzas poblacionales son iguales.

También se obtiene $P[T > 3.198]=0.004$ que nos lleva a rechazar H_0 en una prueba unilateral.

10.7 Prueba de la diferencia entre dos medias con observaciones aparejadas

Sea $(X_1,Y_1), (X_2,Y_2),\ldots, (X_n,Y_n)$ una muestra aleatoria de n datos aparejados, donde las muestras $X_1, X_2,\ldots, X_n$, e $Y_1, Y_2,\ldots, Y_n$, **correlacionadas**, son seleccionadas respectivamente de dos poblaciones normales $X \sim N(\mu_1, \sigma_1^2)$ y $Y \sim N(\mu_2, \sigma_2^2)$.

Podemos concebir estas n diferencias:

$$D_1 = X_1 - Y_1,\ \ D_2 = X_2 - Y_2, \ldots,\ D_n = X_n - Y_n$$

como una muestra aleatoria seleccionada de una población de diferencias $D = X - Y$ cuya distribución es normal $N(\mu_D, \sigma_D^2)$, con media $\mu_D = \mu_1 - \mu_2$ y varianza $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\operatorname{cov}(X, Y)$.

Si σ_D^2 es conocida, la estadística $\overline{D}$, media de las diferencias, tiene distribución normal $N(\mu_D, \sigma_D^2/n)$. Consecuentemente la estadística:

$$Z = \frac{\overline{D} - (\mu_1 - \mu_2)}{\sigma_D/\sqrt{n}}$$

tiene distribución normal $N(0,1)$. Esta estadística Z se utiliza en la prueba de dos medias correlacionadas cuando la varianza σ_D^2 es conocida.

Por otro lado, si σ_D^2 *es desconocida*, entonces, la estadística,

$$T = \frac{\overline{D} - (\mu_1 - \mu_2)}{\hat{S}_D / \sqrt{n}}$$

tiene distribución t-student con $n-1$ grados de libertad, en la que

$$\hat{S}_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}.$$

Esta estadística T se utiliza en la prueba de dos medias correlacionadas cuando la varianza σ_D^2 es desconocida.

Si $H_0 : \mu_D = \mu_1 - \mu_2 = 0$ se supone verdadera, la estadística es:

$$T = \frac{\overline{D}}{\hat{S}_D / \sqrt{n}} \sim t(n-1)$$

Su valor $t_k = \dfrac{\bar{d}}{\hat{s}_d / \sqrt{n}}$ con $\hat{s}_d = \sqrt{\dfrac{\sum_{i=1}^{5}(d_i - \bar{d})^2}{n-1}} = \sqrt{\dfrac{\sum d^2 - n(\bar{d})^2}{n-1}}$ que resulta de una muestra aleatoria correlacionada, se utiliza para probar $H_0 : \mu_D = 0$ contra cualquiera de las alternativas; $H_1 : \mu_D \neq 0$ ó $H_1 : \mu_D > 0$ ó $H_1 : \mu_D < 0$

Las regiones críticas y las reglas de decisión de esta prueba *t*, son similares a los de la sección 10.3

EJEMPLO 10.12.

Un administrador está probando la posibilidad de usar un nuevo paquete de computación. Cambiará de paquete si hay prueba que el nuevo usa menos tiempo que el antiguo al procesar determinada tarea. A fin de tomar una decisión se selecciona una muestra aleatoria de 7 operadoras y se registra el tiempo de procesamiento en segundos con ambos paquetes tal como se da en la tabla que sigue. A partir de estos datos, ¿se cambiará el paquete de computo antiguo por el nuevo?. Use un nivel de significación del 5% y haga las suposiciones necesarias.

SOLUCION.

1. *Hipótesis*: $H_0 : \mu_D = 0$ contra $H_1 : \mu_D > 0$.

2. *Nivel de significación* : $\alpha = 0.05$.
3. *Estadística de la prueba*: Se suponen poblaciones normales. Si la hipótesis nula es verdadera, entonces para $n = 7$, la estadística es

$$T = \frac{\overline{D}}{\hat{S}_D / \sqrt{n}}$$

que se distribuye según una t-Student con $n - 1 = 6$ grados de libertad.

Operadora	Paquete antiguo	Paquete nuevo	Diferencia d_i	d_i^2
1	10	6	4	16
2	10	10	0	0
3	15	12	3	9
4	12	8	4	16
5	16	15	1	1
6	14	16	–2	4
7	16	12	4	16
Total			14	62

4. *Región crítica*: Para $\alpha = 0.05$ y la alternativa unilateral cola derecha en la distribución $t(6)$ se halla el valor crítico $t_{1-\alpha,\, n-1} = t_{0.95,\, 6} = 1.943$. La región crítica en la variación de T es el intervalo:

$$R.C. = \{T > 1.943\}$$

5. *Cálculos*: De los datos se tiene: $n = 7$, $\overline{d} = \frac{14}{7} = 2$,

$$\hat{s}_d = \sqrt{\frac{\sum d^2 - n(\overline{d})^2}{n-1}} = \sqrt{\frac{62 - 7(2)^2}{6}} = 2.38,$$

$$\text{error estándar: } \hat{s}_d / \sqrt{n} = 0.8996$$

$$t_k = \frac{\overline{d}}{\hat{s}_d / \sqrt{n}} = \frac{2}{0.8996} = 2.22.$$

6. *Decisión*: $t_k = 2.22 \in R.C.$, debemos rechazar H_0.

NOTA. Con el paquete estadístico *MCEST* se obtiene la probabilidad:

$$P = P[T > 2.22] = 0.033.$$

10.8 Prueba de hipótesis acerca de proporciones

10.8.1 Una sola proporción

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n escogida de una población Bernoullí $B(1,p)$, donde el parámetro desconocido p es la proporción de éxitos en la población, y sea la estadística,

$$\overline{P} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{X}{n}$$

la *proporción de éxitos en la muestra*, siendo X el número de éxitos en la muestra

La estadística X tiene distribución exactamente binomial $B(n, p)$.

Si n es suficientemente grande $(n \geq 30)$, la estadística

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\overline{P} - p}{\sqrt{p(1-p)/n}}$$

tiene distribución aproximadamente normal $N(0,1)$.

Si se supone verdadera la hipótesis nula $H_0 : p = p_0$, entonces, la distribución muestral de X es exactamente binomial $B(n, p_0)$, y la de la variable aleatoria

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

es aproximadamente normal $N(0,1)$.

La estadística

$$z_k = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

valor calculado a partir de una muestra aleatoria de tamaño n se utiliza para probar $H_0 : p = p_0$, contra una alternativa unilateral o bilateral.

Las regiones críticas y las reglas de decisión de esta prueba Z son similares a los de la sección 10.2

1) **Prueba bilateral.** Si la prueba es de $H_0 : p = p_0$ contra $H_1 : p \neq p_0$, la región crítica en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha/2} \quad \text{o} \quad Z > z_{1-\alpha/2}\}$$

2) **Prueba unilateral cola derecha.** Si la prueba es de $H_0 : p = p_0$ contra $H_1 : p > p_0$, la región crítica en los valores de Z es el intervalo:

$$R.C. = \{Z > z_{1-\alpha}\}$$

3) **Prueba unilateral cola izquierda.** Si la prueba es de $H_0 : p = p_0$ contra $H_1 : p < p_0$, la región crítica en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha}\}$$

NOTA. Al sustituir la variable $Z = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}} = \dfrac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$ en $R.C.$ se obtienen las regiones críticas respectivas en los valores de las variables X y $\overline{P}$ respectivamente.

NOTA. (Muestras pequeñas) Sea x la cantidad de éxitos en una muestra aleatoria pequeña de tamaño n (n<30)

Prueba bilateral. Si $x < np_0$ se calcula

$$P = P[X \leq x \text{ cuando } p = p_0] = \sum_{k=0}^{x} C_k^n \, p_0^k (1-p_0)^{n-k}$$

y si $x > np_0$ se calcula

$$P = P[X \geq x \text{ cuando } p = p_0] = \sum_{k=x}^{n} C_k^n \, p_0^k (1-p_0)^{n-k}.$$

Se rechaza $H_0 : p = p_0$ si $P \leq \alpha/2$. En caso contrario no se rechaza H_0.

Prueba unilateral cola derecha. Se calcula

$$P = P[X \geq x \text{ cuando } p = p_0] = \sum_{k=x}^{n} C_k^n \, p_0^k (1-p_0)^{n-k}$$

y se rechaza $H_0 : p = p_0$ si el valor de P es menor o igual que el nivel de significación α.

Prueba unilateral cola izquierda. Se calcula

$$P = P[X \le x \text{ cuando } p = p_0] = \sum_{k=0}^{x} C_k^n \, p_0^k (1-p_0)^{n-k}$$

y se rechaza $H_0 : p = p_0$, si el valor de P es menor o igual que el nivel de significación α.

EJEMPLO 10.13.

Un fabricante afirma que el 30% de todos los consumidores prefiere su producto. Con el fin de evaluar esta afirmación se tomó una muestra aleatoria de 400 consumidores y se encontró que 100 de ellos prefieren dicho producto.

a) ¿Es ésta, suficiente evidencia para inferir que el porcentaje de preferencia del producto no es 30%?. Utilice el nivel de significación del 1%.

b) Calcular la probabilidad de tomar la decisión errada de aceptar la afirmación del fabricante cuando la verdadera proporción poblacional de aceptación del producto es 20%.

SOLUCION.

a) Sea p el porcentaje poblacional de preferencia del producto.

1. *Hipótesis*: $H_0 : p = 0.30$ contra $H_1 : p \neq 0.30$

2. *Nivel de significación* $\alpha = 0.01$.

3. *Estadística*: Si $H_0 : p = 0.30$, es verdadera, y n grande la estadística:

$$Z = \frac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\overline{P} - 0.3}{\sqrt{0.3(0.7)/n}}$$

que tiene distribución aproximadamente normal $N(0,1)$.

4. *Región crítica*: Para $\alpha = 0.01$ y una alternativa bilateral, en la distribución de Z se encuentra el valor crítico $z_{0.995} = 2.575$. Luego, la región crítica es el intervalo:

$$R.C. = \{Z < -2.575 \ \text{ o } \ Z > 2.575\}.$$

5. *Cálculo*: $n = 400$, $x = 100$, $\overline{p} = \dfrac{x}{n} = \dfrac{100}{400} = 0.25$

$$ES = \text{Error estándar} = \sqrt{p_0(1-p_0)/n} = \sqrt{0.3 \times 0.7 / 400} = 0.0229$$

$$z_k = \frac{\overline{p} - p_0}{ES} = \frac{0.25 - 0.3}{0.0229} = -2.18$$

6. *Decisión*: Como $z_k = -2.18 \notin R.C.$, no deberíamos rechazar H_0, y concluimos que el fabricante tiene la razón.

b) Al sustituir en $R.C.$ la variable

$$Z = \frac{\overline{P} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\overline{P} - 0.3}{0.0229},$$

se obtiene la región crítica en la variable $\overline{P}$:

$$R.C. = \{\overline{P} < 0.241 \text{ o } \overline{P} > 0.359\}$$

Luego, la probabilidad de aceptar $H_0 : p = 0.30$ cuando realmente $p = 0.20$ (error tipo II) es igual a:

$$\beta = P[\text{aceptar } H_0 / p = 0.20] = P[0.241 \leq \overline{P} \leq 0.359 / p = 0.20]$$

$$\beta = P\left[\frac{0.241 - 0.20}{0.02} \leq \frac{\overline{p} - 0.20}{\sqrt{(0.2 \times 0.8)/400}} \leq \frac{0.359 - 0.20}{0.02}\right]$$

$$\beta = P[2.05 \leq Z \leq 7.95] = 0.0202$$

EJEMPLO 10.14

Se afirma que cierto medicamento que se prescribe para aliviar determinada enfermedad es efectivo en más del 80% de los casos. Al parecer esta afirmación es exagerada por lo que se suministra tal medicamento a una muestra aleatoria de 15 pacientes resultando que 13 de ellos han experimentado alivio, ¿es ésta suficiente evidencia para concluir que realmente el medicamento es efectivo en más del 80% de los casos al nivel de significación del 5%?.

SOLUCION .

Sea X el número de pacientes que se sanan en $n = 15$ casos. Entonces $X \sim B(15, p)$ donde p es el porcentaje de pacientes que se sanan en la población de todos los pacientes que sufren la enfermedad.

1. *Hipótesis*: $H_0 : p = 0.80$ contra $H_1 : p > 0.80$.
2. *Nivel de significación*: $\alpha = 0.05$.
3. *Estadística*: Si la hipótesis nula es cierta, la variable X tiene distribución binomial con $n = 15$ y $p = 0.8$.

4. *Región crítica*: Se rechazará H_0 en favor de H_1 si el valor de $P = P[X \geq 13$ cuando $p = 0.80]$ es menor que $\alpha = 0.05$
5. *Cálculo*:

$$P = P[X \geq 13/p = 0.80] = \sum_{k=13}^{15} C_k^{15} (0.8)^k (0.2)^{15-k} = 0.3970.$$

6. *Decisión*: Dado que $P = 0.3970 > \alpha = 0.05$, no se debe rechazar H_0.

10.8.2 Dos proporciones con observaciones independientes

Sean X_1 y X_2 el número de éxitos en dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas respectivamente de dos poblaciones de Bernoullí $B(1, p_1)$ y $B(1, p_2)$, donde los parámetros desconocidos p_1 y p_2 son las proporciones de éxitos poblacionales respectivos.

Sean además las proporciones de éxitos muestrales respectivas:

$$\overline{P}_1 = \frac{X_1}{n_1} \quad \text{y} \quad \overline{P}_2 = \frac{X_2}{n_2}$$

Para n_1 y n_2 suficientemente grandes ($n_1 \geq 30$ y $n_2 \geq 30$), la variable aleatoria

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}.$$

tiene distribución aproximadamente normal $N(0,1)$.

Si $H_0 : p_1 = p_2$ se supone verdadera, la estadística es;

$$Z = \frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\dfrac{p_c(1-p_c)}{n_1} + \dfrac{p_c(1-p_c)}{n_2}}} \cong N(0,1)$$

donde, p_c **es el valor común de los parámetros** p_1 y p_2 **cuya estimación insesgada** (probar!) es:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \overline{p}_1 + n_2 \overline{p}_2}{n_1 + n_2}$$

La estadística

$$z_k = \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

valor que resulta de dos muestras aleatorias, se utiliza para probar la hipótesis nula $H_0 : p_1 = p_2$, contra una alternativa unilateral o bilateral.

Las regiones críticas y las reglas de decisión de esta prueba Z son similares a los de la sección 10.2

1) **Prueba bilateral.** Si la prueba es de $H_0 : p_1 = p_2$ contra $H_1 : p_1 \neq p_2$, la región crítica en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha/2} \quad \text{o} \quad Z > z_{1-\alpha/2}\}$$

2) **Prueba unilateral cola derecha.** Si la prueba es de $H_0 : p_1 = p_2$ contra $H_1 : p_1 > p_2$, la región crítica en los valores de Z es el intervalo:

$$R.C. = \{Z > z_{1-\alpha}\}$$

3) **Prueba unilateral cola izquierda.** Si la prueba es de $H_0 : p_1 = p_2$ contra $H_1 : p_1 < p_2$, la región crítica en los valores de Z es el intervalo:

$$R.C. = \{Z < -z_{1-\alpha}\}$$

EJEMPLO 10.15.

Un patrocinador de un programa especial de televisión afirma que el programa representa un atractivo mayor para los televidentes hombres que para las mujeres, pero, el personal de producción del programa piensa que es igual el porcentaje de televidentes hombres y mujeres que ven el programa especial. Si una muestra aleatoria de 300 hombres y otra de 400 mujeres reveló que 120 hombres y 120 mujeres estaban viendo el programa especial de televisión. ¿puede considerarse significativa la diferencia al nivel $\alpha = 5\%$?

SOLUCION.

Sean p_1 y p_2, respectivamente, las proporciones de hombres y mujeres que ven el programa especial de televisión.

1. Hipótesis $H_0 : p_1 = p_2$ contra $H_1 : p_1 > p_2$.
2. Nivel de significación: $\alpha = 0.05$.
3. Estadística. Si $H_0 : p_1 = p_2$ es verdadera y las muestras son grandes, la estadística es:

$$Z = \frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}}$$

que tiene distribución aproximadamente normal $N(0,1)$.

4. Región crítica. Para $\alpha = 0.05$ y una prueba unilateral cola a la derecha, la región crítica es:

$$R.C. = \{Z > 1.645\}$$

5. Cálculo. Los datos de la muestra dan

Hombres	mujeres
$n_1 = 300$	$n_2 = 400$
$p_1 = 120$	$p_2 = 120$

$$\overline{p}_1 = \frac{120}{300} = 0.4 , \qquad \overline{p}_2 = \frac{120}{400} = 0.3$$

$$\hat{p} = \frac{120+120}{300+400} = 0.34$$

$$ES\text{=Error estándar} = \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} = \sqrt{\frac{0.34\times 0.66}{300} + \frac{0.34\times 0.66}{400}} = 0.03618$$

$$z_k = \frac{\overline{p}_1 - \overline{p}_2}{ES} = \frac{0.4-0.3}{0.03618} = 2.764$$

6. Decisión: Como $z_k = 2.764 \in R.C.$, deberíamos rechazar H_0.

EJERCICIOS

Error tipo I y error tipo II

1. Halle la región de rechazo de la prueba de la hipótesis nula: $H_0: \mu = 20$, contra $H_1: \mu \neq 20$ correspondiente a la media de una población con varianza $\sigma^2 = 16$. Suponga $\alpha = 0.05$ y $n \geq 30$.

2. Halle la región de rechazo de la prueba de la hipótesis $H_0: \mu = 50$ contra $H_1: \mu < 50$ donde μ es la media de una población normal con varianza σ^2 desconocida. Se dan $n = 9$ y $\alpha = 0.01$.

3. De una población normal con media μ y varianza 256 se extraen muestras aleatorias de tamaño 16. Si para comprobar la hipótesis nula $H_0: \mu = 60$ contra $H_1: \mu = 75$, se utiliza $\alpha = 0.05$, hallar β.
Rp. $\beta = 0.0174$.

4.. Para comprobar la hipótesis nula $H_0: \mu = 50$ contra $H_1: \mu \neq 50$, donde μ es la media de una población con desviación estándar 18, se extrae una muestra aleatoria de tamaño 36. Si se utiliza la región de aceptación: $43 \leq \overline{X} \leq 57$,
a) Hallar α, b) Hallar β si realmente $\mu = 63$.
Rp. a) $\alpha = 0.0198$., b) $\beta = 0.0228$.

5. Una urna contiene 7 bolas de las cuales r son rojas y el resto azules. Para probar la hipótesis $H_0: r = 2$ contra $H_1: r > 2$, se extraen 2 bolas de una en una sin reemplazo y se rechaza H_0 si ambas bolas son rojas.
a) Hallar α de la prueba
b) Hallar β cuando $r = 3$.
Rp. a) α=P[ambas rojas / r=2]=0.0476, b) β=P[ambas NO rojas / r=3]=0.857

6. El número de cheques sin fondo que recibe por día un banco es una variable aleatoria X que tiene distribución de Poisson con parámetro λ. Si para comprobar la hipótesis $H_0: \lambda = 5$, versus $H_1: \lambda = 3$, se usa la región de crítica $RC = \{X < 2\}$,
a) hallar la probabilidad de error tipo I.
b) hallar la probabilidad de error tipo II.
Rp. a) $\alpha = P[X \leq 1/\lambda = 5] = 0.04$, b) $\beta = P[X \geq 2/\lambda = 3] = 0.8$..

7. Se ha determinado que el tiempo de operación de un sistema entre una falla y la siguiente tiene distribución exponencial con una media de 10 horas. Se teme que el tiempo medio entre dos fallas consecutivas ha bajado a 8 horas. Para comprobar estas hipótesis cada cierto tiempo se hace una medición del tiempo X entre dos fallas consecutivas y se decide que si $X < 9$ horas se acepta que el tiempo medio entre fallas ha disminuido a 8 horas, de otro modo se acepta que el tiempo medio entre dos fallas consecutivas es 10.

a) Calcule el nivel de significación de la prueba.
b) Calcule la probabilidad de error tipo II.

Rp. a) H_0:μ=10, H_1:μ=8, α=P[X<9 / μ=10]=1−$e^{-9/10}$ ≅ 0.593, b) β=P[X≥9/μ=8]=$e^{-9/8}$≅ 0.325

Una media

8. Un productor de cápsulas de uña de gato afirma que la demanda promedio de su producto en el mercado es de 1000 cápsulas diarias. Sin embargo, un estudio de la demanda de su producto en 36 días aleatorios da una media y una desviación estándar de 850 y 360 cápsulas diarias respectivamente. ¿es esto suficiente evidencia para contradecir la afirmación de este productor?. Utilice el nivel de significación $\alpha = 1.5\%$ en una prueba unilateral.

Rp. H_0 : μ ≥1,000, H_1 : μ <1,000, z_k=−2.5, RC={Z< −2.17}, se rechaza H_0.

9. La duración de cierta marca de baterías es una variable aleatoria cuya distribución se supone normal. Se estima que su duración media es de 500 horas y que el 95% del total duran entre 480.4 y 519.6 horas. Si en una muestra aleatoria de 9 de tales baterías se encuentra que la duración media es 495 horas. ¿es esto evidencia para concluir al nivel de significación del 5% que la duración media de todas esas baterías es diferente de 500 horas?.

Rp. H_0 :μ =500, H_1:μ ≠500, σ=10, RA={−1.96≤Z≤ 1.96}, z_k = −1.5. No

10. Se afirma que los fumadores adultos del país consumen en promedio al menos 10 cigarrillos por día. Para comprobar esta afirmación, se escoge una muestra aleatoria de 36 fumadores adultos y se observa X_i el número de cigarrillos que fuman por día, resultando: $\sum_{i=1}^{36} X_i = 324$ y $\sum_{i=1}^{36} X_i^2 = 3231$. Utilizando α=0.01.

a) ¿Parecería esto indicar que el promedio del consumo es menor que 10?
b) Encuentre la probabilidad del error del tipo II de la prueba si el valor real de la media es 8 cigarrillos por día.

Rp. a) H_0:μ ≥10, H_1:μ<10, z_k=−2, RC={Z<−2.33} , se acepta H_0. b) β=0.0475

11. Cierta prueba de ingreso universitario tiene una media de 200 puntos y una desviación estándar de 50 puntos. Si para comprobar el valor de la media se utiliza la región crítica $RC=\{\overline{X} < 190\}$ donde $\overline{X}$ es la media de muestras de tamaño 100,

a) ¿Con qué probabilidad se rechaza $H_0: \mu = 200$ si es verdadera?

b) Hallar el porcentaje de casos en que se acepta H_0 si realmente es $\mu = 180$.

Rp. a) α=0.0228, b) β=0.0228, 228 casos de 10.000.

12. Se afirma que el peso de los alumnos varones de la universidad tiene una media de 68 kg. y una desviación estándar de 3.6 kg. Si para verificar $\mu = 68$ se utiliza la región crítica $RC = \{\overline{X} < 67 \text{ o } \overline{X} > 69\}$ donde $\overline{X}$ es la media de muestras de tamaño 64, ¿en qué porcentaje de casos esta región crítica no detecta una diferencia igual a 2 kg. en el promedio de los pesos y por encima de 68 kg.?.

Rp. β=0.0132 en 132 casos de 10,000.

13. Se cree que el tiempo promedio que utilizan los alumnos del ciclo básico para realizar cierta prueba de aptitud tiene distribución normal cuya media es 15 minutos. Para comprobar la hipótesis respecto a la media se toma una muestra aleatoria de 16 de tales alumnos y se encuentra un promedio de 16 minutos. Realice una prueba unilateral

a) Con el nivel de significación $\alpha = 0.05$, si sabe que $\sigma = 3.2$

b) Con el nivel de significación $\alpha = 0.05$, si $\hat{s} = 3.2$ se calcula de la muestra.

c) Utilizando el método de la probabilidad P, si sabe que $\sigma = 3.2$.

Rp. H_0:$\mu \leq 15$, H_1:$\mu > 15$, a) z_k=1.25, RC={Z>1.645}, b) t_k=1.25, gl=15, RC={T>1.753}, c) P=0.1056..

14. Cierta prueba de inteligencia para estudiantes preuniversitarios tiene una media de 100 puntos. Para verificar el valor de la media se aplicó la prueba a una muestra aleatoria de 36 estudiantes preuniversitarios dando una media de 90 puntos y una desviación estándar de 30 puntos. Si $\alpha = 0.01$, ¿cuál es la probabilidad de rechazar en forma acertada que el promedio de la prueba es 100 puntos cuando realmente es 80 puntos?.

Rp. H_0:μ =100. H_1:μ =80. RC={ $\overline{X}$ <88.35}, 1−β = 0.9525.

15. Un fabricante afirma que el nuevo hilo sintético que produce tiene una resistencia media a la ruptura mayor de 15 kilogramos. Para probar esta hipótesis se escoge una muestra de 36 de tales hilos encontrando una media y una desviación estándar de resistencia a la ruptura de 16 y 3 kg. respectivamente. Utilizando $\alpha = 0.05$,

a) Probar la afirmación del fabricante comparando α con $P = P[\overline{X} > 16]$.

b) Hallar el porcentaje de las veces en que tal muestra nos lleva a rechazar en forma acertada que la resistencia media a la ruptura es igual a 15 kg. cuando realmente es igual a 2 kg. por encima de ello.

Rp. a) H_0:μ =15, H_1:μ >15,RC={ $\overline{X}$ >15.8225}, P=0.0228, se rechaza H_0, b) 1−β=0.9909 en 9909 casos de 10,000.

16. El gerente de ventas de una compañía afirma que sus vendedores venden semanalmente en promedio $1,500.

a) Al nivel de significación del 5% pruebe la hipótesis del gerente versus la hipótesis del presidente de los vendedores que afirma que el promedio de las ventas semanales es mayor, si una muestra de 36 vendedores ha dado una media igual a $1510 y una varianza igual a 900$^{2} en una semana.

b) ¿Con qué probabilidad la prueba anterior **no detecta** la diferencia igual a 20$ diarios en el promedio de ventas por día y por encima de lo que se indica en la hipótesis nula?.

Rp. a) H_0:μ=1,500, H_1:μ>1,500, RC={ $\overline{X}$ >1508.225} se rechaza H_0. b) β=0.0091.

17. Los sacos de café que recibe un exportador deben tener un peso promedio de 100 kilogramos. Un inspector tomó una muestra de 50 sacos de un lote de 500 sacos de café encontrando una media de 98 Kg. y una desviación estándar de 3 Kg.. Con $\alpha = 0.02$ y mediante una prueba unilateral

a) ¿Es razonable que el exportador rechace el lote de sacos de café?.

b) ¿Con qué probabilidad esta prueba de hipótesis **detecta** la diferencia igual a 2 Kg. en el peso promedio del lote y por debajo de lo que se requiere para exportar?.

Rp. Población finita a) H_0:μ =100. RC= {$\overline{X}$ <99.174}, se rechaza H_0.

b) P[$\overline{X}$ < 99.174/μ=98]=0.9982.

18. Un fabricante está considerando la adquisición de un nuevo equipo para enlatar conservas de palmito y especifica que el contenido promedio debe ser 300 gramos por lata. Un agente de compras hace una visita a la compañía donde está instalado el equipo y observa que una muestra aleatoria de 10 latas de palmito ha dado los siguientes pesos en gramos.

Pesos	296	297	298	299	300	301	302
# de latas	2	2	2	1	1	1	1

Y encuentra además que provienen de una población normal. Probar la hipótesis nula que la media poblacional es 300 gramos contra una alternativa bilateral.

a) Utilizando un nivel de significación del 5%,

b) Por el método de la probabilidad P. (Utilice un paquete de computo)

Rp. H_0 : μ =300, H_1:$\mu \neq$300, error estándar=0.653, t_k=−2..45, gl=9. a) RA={−2.262≤T≤ 2.262}, se rechaza H_0. b) el *MCEST* da P=P[T<−2.45]=0.0177, signif. bilateral=0.0354.

19. El promedio de nicotina que tiene cierta marca de cigarrillos es 10 miligramos por cigarrillo. El fabricante afirma que un nuevo proceso de fabricación reducirá el promedio de la nicotina por cigarrillo. Para comprobar esta hipótesis se toma una muestra aleatoria de 9 cigarrillos fabricados con el nuevo proceso y se encuentran las siguientes cantidades de nicotina en miligramos:

9, 9.2, 8.5, 8.7, 9, 8.8, 9.2, 9.4, 9.2

y se encuentra además que provienen de una población normal. ¿Es razonable aceptar la afirmación del fabricante?

a) Utilice un nivel de significación del 1%,
b) Por el método de la probabilidad P. (Utilice un paquete de computo)

Rp. a) H_0:$\mu \geq 160$, H_1:$\mu < 160$., $\bar{x} = 9$, $\hat{s} = 0.2872$, ES=0.0957, $t_k = -10.4$, $RC = \{T < -2.896\}$, se rechaza H_0. b) el *MCEST* da $P = P[T < -10.4] = 0.000$.

20. Se sabe que las ventas diarias de una compañía tienen distribución normal con una desviación estándar de S/300. El gerente de la compañía afirma que en promedio las ventas diarias de la compañía es por lo menos S/.2,277 se trata de probar, con $\alpha = 0.004$; si la afirmación del gerente es verdadera; para esto se tomará una muestra aleatoria de tamaño n . Hallar n y la región crítica de la prueba sabiendo que si la verdadera media es 1800. entonces la probabilidad de error tipo II sería igual a 0.017 Rp. $n = 9$. $RC = \{\bar{X} < 2012\}$

21. Se afirma que el ingreso promedio mensual de un sector de informales es $400. Diseñe una prueba para probar esta hipótesis con un riesgo de 5% de cometer error tipo I y un riesgo de 1% de cometer error tipo II cuando realmente es μ=$460. Suponga que σ=$90.

Rp. $n \cong 36$, $K = 424.675$.

22. La duración de cierto tipo de focos de luz se distribuye normalmente con una media de 400 horas y una desviación estándar de 24 horas. Se está considerando aumentar la duración promedio con un nuevo proceso. Si la duración promedio aumenta 15 horas, este cambio debe detectarse con probabilidad 0.9554. Si no hay cambio, este debe detectarse con probabilidad 0.98. Determine el número de focos que deben probarse y la región crítica.

Rp. a) $n = 16$, b) $RC = \{\bar{X} > 408.2\}$.

Una proporción

23. Se controla la calidad de una muestra aleatoria de 40 piezas producidas por un fabricante. Si se hallaron 4 piezas defectuosas, ¿se debería inferir que el porcentaje de todas las piezas defectuosas es más del 5% al nivel de significación del 5%?.

Rp. H_0: p=0.05. H_1: p>0.05. $RC = \{Z > 1.645\}$, $z_k = 1.45$ se acepta H_0.

24. Una firma va a comercializar un nuevo producto sólo si hay prueba de que al menos el 20% de todos los consumidores lo prefieren. Para probar esa hipótesis se selecciona al azar 200 consumidores. Si se utiliza como región crítica

$\{ X < 30\}$ donde X es el número de consumidores en la muestra que prefieren el producto, calcular la probabilidad de cometer error tipo I

Rp. $\alpha = 0.0314$ con corrección por continuidad.

25. Se afirma que el 20% de todos los electores están a favor de cierto candidato. Para verificar esta hipótesis se escogen 400 electores al azar y si la proporción a favor en la muestra; $\overline{p}$; está entre 16.08% y 23.92% se acepta que la proporción a favor en la población es $p = 20\%$. En caso contrario se acepta que $p \neq 20\%$.

a) ¿Cuál es la probabilidad de cometer error tipo I?.

b) Calcular la probabilidad cometer error tipo II cuando $p = 0.25$.

Rp. a) α =0.05, b) β=0.3085

26. Tradicionalmente el 13% de los conductores de fin de semana conducen bajo los efectos del alcohol. El último fin de semana fueron intervenidos 500 conductores aleatoriamente y 80 de ellos estaban bajo los efectos del alcohol. ¿De esta muestra se puede inferir que el porcentaje poblacional ya no es 13%?. Utilice $\alpha = 5\%$.

Rp. z_k=2., a) $RA.= \{-1.96 \leq Z \leq 1.96\}$, se rechaza H_0: p=0.13

27. El gerente de una tienda afirma que el 80% de los clientes del año pasado, regresarán este año a realizar sus compras. Sin embargo, analizando el mercado, nosotros creemos que dicho gerente ha exagerado. Para probar estas hipótesis se toma una muestra aleatoria de 200 clientes que el año pasado habían comprado en dicha tienda. Si $\alpha = 0.05$ y si la verdadera proporción de clientes que regresan a la tienda es 70%. Calcular β.

Rp. β=0.0495

28. El Director de la bolsa de trabajo de la universidad afirma que el 10% de los egresados de la Universidad consiguen empleo con una remuneración mayor de $3,000 mensuales. Al parecer el porcentaje indicado es optimista. Para comprobar esta afirmación se debe tomar una muestra aleatoria de n egresados. Hallar el tamaño de la muestra y la regla de decisión si se desea que la probabilidad de cometer error tipo I sea 0.2514 y que el riesgo de tomar una decisión equivocada cuando la proporción de egresados con una remuneración mayor de $3,000 sea del 5% con una probabilidad de 0.0853

Rp. $n \cong 100$. $K \cong 0.08$

29. Un legislador desea probar la hipótesis que mas del 65% de sus representados está a favor de cierta legislación laboral que se está presentando en el congreso. Para esto, realiza una consulta a 400 electores seleccionados al azar. Considerando $\alpha = 0.05$.

a) ¿Qué valor como mínimo debe tener la proporción de la muestra, para que a partir de ese valor, la decisión sea aceptar la hipótesis del legislador?.
b) ¿Cuál es la probabilidad de tomar la decisión errada de rechazar la propuesta del legislador cuando en realidad el 70% de los votantes acepta la legislación laboral?

Rp. a) K=0.689. b)$P[\overline{P} \leq 0.689 / p=0.7] = P[Z\leq -0.48]$=0.3156.

30. Se asegura que el 70% de los trabajadores están asegurados bajo el régimen particular de pensiones (AFP). Para probar esta afirmación se toma una muestra de 80 personas que trabajan. Si menos de 52 personas de la muestra están aseguradas en el régimen indicado, se rechaza que el 70% de la población de trabajadores está asegurado en AFP.
a) ¿Cuál es el nivel de significación de la prueba?
b) ¿Se podría decir que la prueba puede detectar una diferencia de 20% por debajo de lo indicado en la hipótesis nula?.

Rp. a) α=0.1635, b) $\beta=P[\overline{P} \geq 0.65 / p = 0.50]=P[Z \geq 2.68]$= 0.0037. La prueba detecta la diferencia el $1-\beta$= 99.63% de las veces.

31. De una lista de 2,000 clientes de un banco comercial se seleccionó una muestra aleatoria para obtener opinión acerca del servicio. En la muestra se halló que 215 no tienen quejas del servicio, 25 tienen quejas y 10 no opinan al respecto. Tradicionalmente el 5% tenían quejas del servicio, sin embargo se cree que ahora este porcentaje aumentó. ¿Cuál es la situación actual si se quiere una probabilidad de 0.007 de cometer error tipo I?.

Rp. Población finita H_0: p=0.05, H_1: p>0.05, $RC=\{Z>2.455\}$, z_k=3.88 , se rechaza H_0.

Varianzas

32. Una muestra aleatoria de 16 sobres de cierto producto cuyos pesos se distribuyen normalmente ha dado una desviación estándar de 0.6 gramos. Utilizando un nivel de significación del 5%, ¿es válido inferir que la varianza de los pesos de tales sobres es mayor que 0.25 gramos^2 ?

Rp. H_0: σ^2=0.25, H_1: σ^2>0.25 , $X\sim\chi^2(15)$, $RC=\{X>25\}$, x_k=21.6, se acepta H_0.

33. Anteriormente la desviación estándar de los pesos de los contenidos de cierto envase era 0,25 onzas, se trata de averiguar si ha habido aumento de dicha variabilidad , para esto se toma una muestra aleatoria de los contenidos de 20 envases encontrándose una desviación estándar de 0,30 onzas. Al nivel de significación del 5%, ¿proporcionan los datos indicios suficientes que indique un aumento significativo de tal variabilidad?. Suponer que dichos pesos están normalmente distribuidos.

Rp. H_0: σ = 0.25, H_1: σ > 0.25, $X\sim\chi^2(19)$, $RC=\{X>30.14\}$, x_k=27.36, se acepta H_0.

34. Se escoge una muestra aleatoria de 13 tiendas y se encuentra que las ventas de la semana de un determinado producto de consumo popular tiene una desviación estándar $\hat{s} = \$6$. Se supone que las ventas del producto tienen una distribución normal. Al nivel de significación del 5%, ¿se podría inferir que la varianza de la población es menor que $40\$^2$?.

Rp. H_0: $\sigma^2=40$, H_1: $\sigma^2<40$, $X \sim \chi^2(12)$, $RC=\{X<5.23\}$, $x_k=10.8$, se acepta H_0.

35. Una de las maneras de medir el grado de satisfacción de los empleados de una misma categoría en cuanto a la política salarial, es a través de las desviaciones estándar de sus salarios. La fábrica A afirma ser más homogénea en la política salarial que la fábrica B. Para verificar esa afirmación, se escoge una muestra aleatoria de 10 empleados no especializados de A, y 9 de B, obteniendo las dispersiones $\hat{s}_A = 10$, $\hat{s}_B = 15$ de salario mínimo. Suponiendo poblaciones normales y con α=0.01, ¿cuál sería su conclusión?

Rp. H_0: $(\sigma_1)^2=(\sigma_2)^2$, H_1: $(\sigma_1)^2<(\sigma_2)^2$, $F \sim F(9,8)$, $RC=\{F<0.183\}$, $f_k=0.44$, se acepta H_0.

36. Los tiempos en minutos para realizar cierta tarea observados en 10 hombres y 10 mujeres fueron:

Hombres: 50, 45, 49, 50, 38, 58, 53, 47, 48, 55
Mujeres: 55, 56, 57, 56, 58, 53, 54, 59, 60, 57

Suponiendo poblaciones normales, ¿se podría concluir que las varianzas poblacionales son diferentes?,

a) Utilizando el nivel de significación del 5%
b) Utilizando la probabilidad P

Rp. H_0: $(\sigma_1)^2=(\sigma_2)^2$, H_1: $(\sigma_1)^2\neq(\sigma_2)^2$, $F \sim F(9,9)$, $f_k=6.496$, a) $RC=\{F<0.248$ o $F>4.03\}$, se rechaza H_0, b) $P=P[F>6.496]=0.0057$, significación bilateral 0.0114, se rechaza H_0.

Diferencia de dos medias

37. Para comparar la aptitud de dos poblaciones de estudiantes preuniversitarios se toman dos muestras aleatorias respectivas de tamaños 20 y 25, dando las medias respectivas de 200 y 205 puntos. Suponga que las dos poblaciones son normales con $\sigma_1 = 8$, y $\sigma_2 = 7$. Al nivel de significación del 1%, ¿se podría concluir que las medias de las dos poblaciones son distintas?.

Rp H_0:$\mu_1=\mu_2$, H_1:$\mu_1\neq\mu_2$ $z_k=-2.2$, $RC=\{Z<-2.575$ o $Z>2.575\}$ se acepta H_0

38. Un inversionista está por decidir entre dos provincias para abrir un centro comercial. Para esto debe probar la hipótesis de que hay diferencia en el promedio de ingresos familiares de las dos provincias. Si una muestra de 300 hogares de la provincia 1 da $\bar{x}_1 = \$400$ y $\hat{s}_1 = \$90$ y otra muestra de 400 hogares de la provincia 2 da $\bar{x}_2 = \$420$ y $\hat{s}_2 = \$120$. ¿Se puede inferir que las

dos medias poblacionales son diferentes?, si es así, ¿en cual de las provincias debería abrir la sucursal?. Utilice $\alpha = 0.05$

Rp. $z_k=-2.52$, $R.C=\{Z<-1.96 \text{ o } Z>1.96\}$, $\mu_1 \neq \mu_2$, además $\mu_1<\mu_2$.

39. Se quiere determinar la diferencia entre los promedios de tiempos (en minutos) que utilizan los hombres y las mujeres para realizar determinada tarea. Con este fin se escogen 16 hombres y 16 mujeres al azar resultando los tiempos promedios respectivos 40 y 35 minutos, y desviaciones estándar respectivos 9 y 8 minutos. Suponga que las poblaciones de ambos tiempos son independientes y que se distribuyen normalmente con varianzas iguales. Al nivel de significación del 1% ¿es el tiempo promedio de hombres mayor al tiempo promedio de mujeres?.

Rp. gl=30, error estándar=3.01, $t_k=1.66$, $RC=\{T>2.457\}$, No, $\mu_1=\mu_2$.

40. Una compañía debe decidir cuál de dos tipos de componente electrónica A o B va a adquirir. Hace una prueba de 5 componentes escogidos al azar para cada marca, resultando $\bar{x}_1 = 8000$ y $\hat{s}_1 = 2500$ horas para A y $\bar{x}_2 = 7000$ y $\hat{s}_2 = 800$ horas para B. Suponga poblaciones normales con varianzas diferentes. Pruebe la hipótesis nula que los rendimientos medios son iguales contra la alternativa de que A rinde más que B. Use $\alpha = 0.05$.

Rp. gl=5, error estándar=1173.88, $t_k=0.85$, $RC=\{T>2.015\}$, $\mu_1=\mu_2$.

41. Se afirma que una nueva dieta reduce el peso de una persona en 5 kilogramos promedio en un periodo de un mes. Se registran los pesos de 12 mujeres que siguieron esta dieta antes y después del período resultando $\bar{x}_1 = 62$ y $\bar{x}_2 = 58$ kilogramos y una desviación estándar de las diferencias de pesos $\hat{s}_d = 5$kg. Utilizando un nivel de significación del 5%, Verifique la afirmación contra la alternativa

a) La diferencia de peso es diferente de 5kg.
b) La diferencia de peso es mayor de 5kg.

Suponga que la diferencia de los pesos tiene distribución normal.

Rp. $\bar{d} = 4$, $s_d = 5$, error estándar=1.44, gl=11, $t_K=2.77$, a) $RA=\{-2.201 \leq T \leq 2.201\}$, b) $RC=\{T>1.796\}$, se rechaza medias iguales en ambos casos.

42. Un agente de compras de una compañía se vio confrontado con dos marcas de computadoras para su adquisición. Se le permitió probar ambas marcas asignando una misma tarea a 50 máquinas de cada marca, resultando las medias respectivas 55 y 50 minutos. Suponga las dos poblaciones tienen varianza homogénea igual a 100. Para $\alpha = 0.05$

a) ¿Excede el tiempo promedio de la marca 1 al de la marca 2 en al menos 9 minutos?.

b) Hallar la potencia de la prueba cuando la diferencia real entre promedios de tiempo de marca 1 menos marca 2 sea 3 minutos.

c) ¿Qué tan grande debe ser la muestra si la potencia de la prueba es 0.95, cuando la diferencia real entre promedios de tiempo marca 1 menos marca 2 es 3 minutos

Rp. a) H_0:$\mu_1-\mu_2\geq 9$, H_1:$\mu_1-\mu_2<9$, E.S=2, $z_k=-2$, $RC=\{Z<-1.645\}$ se rechaza H_0.,
b) $1-\beta=0.9131$, c) $n=105$

Utilizando un paquete de computo (por ejemplo *MCEST*) resolver:

43. Para comparar los promedios de los tiempos en minutos que emplean dos máquinas 1 y 2 en producir un tipo de objeto, se registra el tiempo de 9 y 8 objetos al azar producidos por las máquinas 1 y 2 respectivamente dando los siguientes resultados:

Máquina 1: 12, 28, 10, 25, 24, 19, 22, 33, 17.
Máquina 2: 16, 20, 16, 20, 16, 17, 15, 21.

Al nivel de significación del 5%, ¿confirman estos datos que los tiempos promedios de las dos máquinas son diferentes?. Suponga que los tiempos en ambos casos se distribuyen normalmente

Rp. F=10.186, gl:8,7, sigf.bilat=0.007, se acepta $(\sigma_1)^2\neq(\sigma_2)^2$, t_k=1.337, g.l.=10, ES=2.608,
sigf.bilat=0.209, se acepta $\mu_1=\mu_2$.

44. Una compañía de transporte terrestre de pasajeros está por decidir si comprar una marca A o una marca B de llantas para su flota de ómnibuses. Se prueban 9 llantas escogidas al azar de cada una de las marcas resultando los siguientes rendimientos en kilómetros:

Marca A: 32000, 30000, 33000, 31000, 32000, 35000, 34000, 35000, 31000
Marca B: 35000, 37000, 36000, 38000, 37000, 39000, 32000, 33000, 40000

Suponiendo poblaciones normales y con $\alpha = 0.01$

a) ¿Se podría concluir que las varianzas son iguales?

b) ¿Se puede concluir que las dos marcas rinden igual?. Si no es así, ¿qué marca rinde más?.

Rp. a) F=2.136, $RA=\{0.13\leq F\leq 7.50\}$, sigf.bil=0.3037 se acepta $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=−3.54, g.l.=16,
ES=1068.632, $RA=\{-2.921\leq T\leq 2.921\}$, sigf.bil=0.003, se rechaza $\mu_1=\mu_2$.

45.. Una empresa envasa en frascos de 300 gramos uno de sus productos que tiene dos componentes A y B, en iguales cantidades promedio. Una muestra aleatoria de 10 frascos ha dado los siguientes porcentajes de la componente A

48%, 52%, 49%, 55%, 62%, 51%, 53%, 54%, 55%, 56%

Suponiendo que los contenidos A y B tienen distribución normal y $\alpha = 0.05$.

a) Determine si las varianzas de los contenidos de ambas componentes son iguales.

b) ¿Son diferentes los promedios de los contenidos de ambas componentes?.

Rp. a) A: 144, 156, 147, 165, 186, 153, 159, 162, 165, 168, B: 156, 144, 153, 135, 114, 147, 141, 138, 135, 132, F=1, gl:9,9, sigf.bil=1.00 se acepta $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=3.934, g.l=18, ES=5.34, sigf.bil=0.001, se rechaza $\mu_1=\mu_2$.

46. Para un bien de consumo popular del fabricante A, una muestra de 10 tiendas ha dado las siguientes ventas (en dólares):

32000, 30000, 33000, 31000, 32000, 35000, 34000, 35000, 31000, 33000

Para el mismo bien de consumo popular del fabricante B, una muestra de 11 tiendas ha dado las siguientes ventas:

35000, 32000, 36000, 38000, 37000, 39000, 38000, 40000, 42000, 45000, 44000

Suponiendo poblaciones normales y $\alpha = 0.05$

a) ¿Se podría concluir que las varianzas son iguales?

b) ¿Se podría concluir que son iguales los promedios de ventas de los dos productos?. Si no es así, ¿qué producto se vende más?.

Rp. a) F=5.12, gl:10,9, sigf.bil=0.022 se rechaza $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=−4.758, g.l=14, ES=1287.875, sigf. bil=0.003, sigf.unil=0.0015, se acepta $\mu_1<\mu_2$ B vende más.

47. Una compañía debe decidir entre dos tipos de fluorescentes para todos los ambientes de su local. Por cuestiones de precio la compañía desearía comprar los fluorescentes de marca A, a menos que haya evidencia que la marca B tenga mayor duración que la marca A. Se prueban 8 fluorescentes de cada marca y se obtienen las siguientes duraciones en horas:

Marca A: 1500, 1700, 1600, 1800, 1700, 1900, 1200, 1300
Marca B: 1200, 1000, 1300, 1100, 1200, 1500, 1400, 1500

Suponiendo normalidad y utilizando un nivel de significación del 5%,

a) ¿Se podría concluir que las varianzas poblacionales son iguales?

b) ¿Qué marca debe comprar la empresa?.

Rp. F=1.739, gl:7,7, sigf.bil=0.482 se acepta $(\sigma_1)^2=(\sigma_2)^2$, b) t_k=2.915, g.l=14, ES=107.217, sigf.bil=0.11, sigf.unil=0.005, se acepta $\mu_1>\mu_2$ adquirir marca A.

48. Un encargado de compras de una compañía tiene que escoger entre dos tipos de máquinas A y B, para hacer cierta operación. Se le permitió probar ambas máquinas durante un periodo de prueba para lo cual se asignan 10 tareas similares a cada una de las dos máquinas y se obtienen los siguientes tiempos en segundos:

Máquina A: 55, 56, 57, 56, 58, 53, 54, 59, 60, 57
Máquina B: 50, 45, 49, 50, 38, 58, 53, 47, 48, 55

Suponiendo normalidad y utilizando un nivel de significación del 5%,

a) ¿Se podría concluir que las varianzas poblacionales son iguales?

b) ¿Qué tipo de máquina debería comprar la empresa?.

Rp. F=6.946, gl:9,9, sigf.bil=0.011 se acepta $(\sigma_1)^2 \neq (\sigma_2)^2$, b) t_k=3.83, g.l=12, ES=1.88, sigf.bil=0.002, sigf.unil=0.001, se acepta $\mu_1 > \mu_2$ adquirir máquina B.

49. Una compañía de transporte interprovincial debe decidir si compra la marca A o la marca B de neumáticos para su flota de ómnibus. Para estimar la diferencia entre las dos marcas asigna un neumático de cada marca a las ruedas delanteras de 12 ómnibus y se registran en miles de kilómetros las siguientes distancias:

Omnibus	1	2	3	4	5	6	7	8	9	10	11	12
Marca A	50	47	38	44	35	36	44	48	46	48	49	51
Marca B	45	43	30	39	35	31	42	44	37	46	48	52

Utilizando un nivel de significación del 5%, ¿se puede concluir que los promedios de rendimiento son iguales en ambas marcas con una prueba bilateral?. Suponga que las diferencias de las distancias se distribuyen en forma normal.

Rp. $\bar{d} = 3.67$, $s_d = 3.025$. ES=0.873, gl=11, t_K=4.199, $RA=\{-2.201 \le T \le 2.201\}$

Sigf.bil=0.001, se rechaza $\mu_1 = \mu_2$.

Diferencia de dos proporciones

50. Una empresa de estudios de mercado quiere saber si un producto promocionado a nivel nacional lo adquieren los hombres en mayor porcentaje que las mujeres. Si en dos muestras aleatorias independientes de 900 hombres y 800 mujeres se encontró que 270 hombres y 200 mujeres adquieren el producto, ¿cuál es su decisión al nivel $\alpha = 0.004$?.

Rp. $H_0{:}p_1=p_2$, $H_1{:}p_1>p_2$, z_k=2.3, $RC=\{Z>2.65\}$ lo adquieren en igual cantidad.

51. Verificar la afirmación de que la diferencia $p_1 - p_2$ es menor que 5% donde p_1 y p_2 son las proporciones de objetos defectuosos de dos fabricantes A y B, si dos muestras aleatorias independientes de 200 objetos de cada fabricante dan 20 y 12 objetos defectuosos respectivamente para A y B. Use el nivel de significación: 5%.

Rp. $H_0{:}p_1-p_2=0.05$, $H_1{:}p_1-p_2<0.05$, $z_k=-0.37$, $RC=\{Z<-1.645\}$, se acepta H_0.

52. En una muestra de 500 hogares de Trujillo se encuentra que 50 de ellos están viendo vía satélite un programa especial de televisión. En Tarapoto, 28 hogares de una muestra aleatoria de 400 se encuentran viendo el mismo programa especial. ¿Puede rechazarse la suposición del patrocinador de que el porcentaje de hogares que están observando el programa especial es el mismo en las dos ciudades?. Utilice una prueba bilateral y $\alpha = 0.05$.

Rp. $H_0{:}p_1=p_2$, $H_1{:}p_1 \neq p_2$, z_k=1.59, $RA=\{-1.96 \le Z \le 1.96\}$ se acepta H_0.

53. En un estudio de mercado para determinar el rating de los programas de TV del mediodía una muestra aleatoria de 400 hogares de cierta comunidad revela que 80 están sintonizando el programa de TV B, 120 sintonizan el programa G y el resto sintonizan otra cosa. ¿Es la proporción global de televidentes que sintonizan el programa B igual al que sintonizan G?. Utilice $\alpha = 0.01$ y una prueba bilateral.

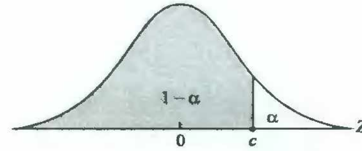
Rp. $H_0: p_1 = p_2$, $H_1: p_1 \neq p_2$, $z_k = -3.27$, $RA = \{-2.575 \leq Z \leq 2.575\}$ se rechaza H_0.

54. Una agencia de publicidad realizó un estudio para comparar la efectividad de un anuncio en la radio en dos distritos. Después de difundir dicho aviso, se realizó una encuesta telefónica con 600 personas seleccionadas al azar, que viven en cada uno de los distritos resultando las proporciones: 20% y 18% respectivamente. Verificar, al nivel de significación del 5%, si son iguales las proporciones de personas que escucharon dicho aviso en los dos distritos mediante una prueba unilateral.

Rp. $z_k = 0.88$, $RC = \{Z > 1.645\}$ se acepta H_0.

TABLA DE LA DISTRIBUCION NORMAL ESTANDAR

La tabla da áreas $1-\alpha$ y valores
$c = Z_{1-\alpha}$, donde, $P[Z \leq c] = 1-\alpha$,
y donde Z tiene distribución normal $N(0,1)$.

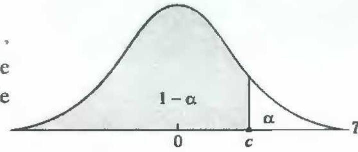


Z	Segundo decimal de Z .00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	1.000	1.000	1.000
4.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

APANDICE B

TABLA DE LA DISTRIBUCION t–Student

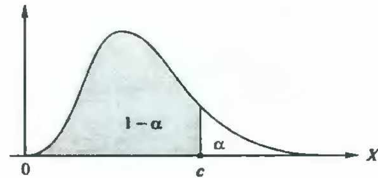
La tabla da áreas $1-\alpha$ y valores $c = t_{1-\alpha,r}$, donde, $P[T \leq c] = 1-\alpha$, y donde T tiene distribución t-Student con r grados de libertad..



	$1-\alpha$							
r	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

TABLA DE LA DISTRIBUCION CHI-CUADRADO

La tabla da áreas $1-\alpha$ y valores $c=\chi^2_{1-\alpha,r}$ tales que $P[X \le c]=1-\alpha$, donde X tiene distribución χ^2 con r grados de libertad.

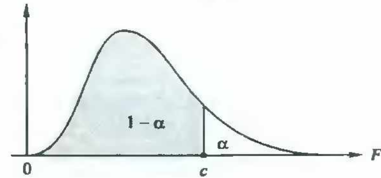


r	$1-\alpha$									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.71	3.84	5.02	6.64	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.82	9.35	11.35	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.65	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.54	20.09	21.96
9	1.74	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.58	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.69	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.27	7.01	8.23	9.39	10.87	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.57	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

APANDICE D

TABLA DE LA DISTRIBUCION DE PROBABILIDADES F

La tabla da áreas $1-\alpha$ y valores $c = F_{1-\alpha, r_1, r_2}$ tales que $P[F \le c] = 1-\alpha$. Donde r_1 y r_2 son los grados de libertad, y donde $F_{\alpha, r_2, r_1} = 1/F_{1-\alpha, r_1, r_2}$.

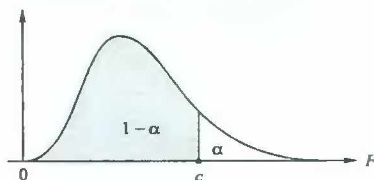


		r_1													
1−α	r_2	1	2	3	4	5	6	7	8	9	10	12	15	20	120
.95	1	161	200	216	225	230	234	237	239	241	242	244	246	248	253
.975		648	800	864	900	922	937	948	957	963	969	977	985	993	1014
.95	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5
.975		38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5
.99		98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5
.995		199	199	199	199	199	199	199	199	199	199	199	199	199	199
.95	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.55
.975		17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	13.9
.99		34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.2
.995		55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.4	43.1	42.8	42.0
.95	4	6.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.66
.975		12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.31
.99		21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.6
.995		31.3	26.9	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.7	20.4	20.2	19.5
.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.40
.975		10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.07
.99		16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.11
.995		22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.4	13.1	12.9	12.3
.95	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.70
.975		8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	4.90
.99		13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	6.97
.995		18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.2	10.0	9.81	9.59	9.00
.95	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.27
.975		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.20
.99		12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.74
.995		16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.19
.95	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	2.97
.975		7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.73
.99		11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	4.95
.995		14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.06
.95	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.75
.975		7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.39
.99		10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.40
.995		13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.30

APANDICE D

TABLA DE LA DISTRIBUCION DE PROBABILIDADES *F*

La tabla da áreas $1-\alpha$ y valores $c = F_{1-\alpha, r_1, r_2}$ tales que $P[F \leq c] = 1 - \alpha$. Donde r_1 y r_2 son los grados de libertad, y donde $F_{\alpha, r_2, r_1} = 1/F_{1-\alpha, r_1, r_2}$.



		r_1													
$1-\alpha$	r_2	1	2	3	4	5	6	7	8	9	10	12	15	20	120
.95	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.58
.975		6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.14
.99		10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.00
.995		12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	4.75
.95	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.34
.975		6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.79
.99		9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.45
.995		11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.01
.95	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.11
.975		6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.46
.99		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	2.96
.995		10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.37
.95	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	1.90
.975		5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.16
.99		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.52
.995		9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	2.81
.95	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.68
.975		5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	1.87
.99		7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.11
.995		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.30
.95	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.47
.975		5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.58
.99		7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	1.73
.995		8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	1.83
.95	120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.35
.975		5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.95	1.82	1.43
.99		6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.53
.995		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.61
.95	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.22
.975		5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.27
.99		6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.32
.995		7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.36

BIBLIOGRAFIA

1. Amón Jesús. Estadística para Psicólogos. Ediciones Pirámide.
2. Calzada Benza José. Estadística General. Editorial Jurídica. Lima.
3. D'Ottone Horacio. Estadística Elemental. CCP. Chile.
4. Fausto I. Toranzos. Estadística. Kapelusz. Buenos Aires.
5. Levin Richard. Estadística para Administradores. Prentice Hall.
6. Lincoln L. Chao. Introducción a la Estadística. CECSA.
7. Lincoyán Portus Goviden. Curso Práctico de Estadística. McGraw-Hill
8. Lind, Mason y Marchal ESTADÍSTICA para Administración y Economía 3era Edición (2000), McGraw-Hill , 575 páginas.
9. Mood Graybill. Introducción a la Teoría de la Estadística. Aguilar.
10. Meyer P.L.. Probabilidad y Aplicaciones Estadísticas. Fondo Educativo Interamericano.
11. Miller y Freund. Probabilidad y Estadística Para Ingenieros. Prentice Hall.
12. Ostle Bernardo. Estadística Aplicada. Limusa.
13. Siegel Sydney. Estadística No Paramétrica. Trillas.
14. Spiegel Murray. Probabilidad y Estadística. McGraw-Hill.
15. Taro Yamane. Estadística. Harla.
16. Walpole y Myer. Probabilidad y Estadística. McGraw-Hill.
17. Ya-Lun Chou. Análisis Estadístico. Interamericana.
18. Manual del SPSS
19. Manual del ESTADÍSTICA .

El uso del método científico para la solución de los problemas de estadística requiere además de los conocimientos básicos la facilidad que le brinda un software adecuado, por ejemplo el :

MCEST : "Métodos Estadísticos Básicos"

El MCEST es un paquete de computo didáctico creado por el autor de este texto, cuyos resultados son compatibles con los obtenidos utilizando los paquetes: SPSS, ESTADÍSTICA, EXCEL, MINITAB, etc.

